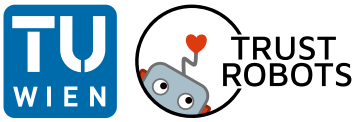
Sabine T. Koeszegi / Markus Vincze (Eds.)

TRUST IN ROBOTS

TU WIEN Academic Press

Sabine T. Koeszegi / Markus Vincze (Eds.)
TRUST IN ROBOTS

This publication emerged from the cooperation between TU Wien Bibliothek and the Doctoral College “Trust in Robots – Trusting Robots”.



Sabine T. Koeszegi / Markus Vincze (Eds.)

TRUST IN ROBOTS

TU WIEN Academic Press

Cite as:
Koeszegi, S. T., & Vincze, M. (Eds.). (2022). *Trust in Robots.* TU Wien Academic Press.
https://doi.org/10.34727/2022/isbn.978-3-85448-052-5

TU Wien Academic Press 2022

c/o TU Wien Bibliothek
TU Wien
Resselgasse 4, 1040 Wien
academicpress@tuwien.ac.at
www.tuwien.at/academicpress

ISBN (print): 978-3-85448-051-8
ISBN (online): 978-3-85448-052-5

Available online: https://doi.org/10.34727/2022/isbn.978-3-85448-052-5

Media proprietor: TU Wien, Karlsplatz 13, 1040 Wien
Publisher: TU Wien Academic Press
Editors (responsible for the content): Sabine T. Koeszegi and Markus Vincze
Production: Facultas Verlags- und Buchhandels AG

Preface

Honesty, responsibility and accountability in all facets of research and university education are the foundations of society's faith in science and technology. These principles serve as the foundation for academic independence and the highest ethical and integrity standards. Therefore, "Technology for People" is the university's mission statement. We want to transform what is technologically feasible into what is desirable from a human-centered perspective. Innovative goods, services and procedures ought to make the world a better place to live in terms of compassion and social responsibility. TU Wien has made significant financial investments in multidisciplinary research carried out in doctoral colleges to address urgent societal concerns to further this purpose and uphold the highest standards of science and ethics. The doctoral college "Trust in Robots", led by Sabine Koeszegi & Markus Vincze, was established in 2018 to understand how we can build disruptive robotic and artificial intelligence (AI) technologies that people trust. Robotics and AI have the potential to help us overcome several problems, including the aging population crisis and climate catastrophe. The doctoral college "Trust in Robots" addresses this area of friction and bargains over the compatibility of technology and moral principles.

"Trust in Robots" has been set up as a transdisciplinary doctoral college in which postgraduate students and professors of various academic disciplines collaborate to understand the same phenomenon from different perspectives. From an institutional standpoint, the College's setup has been difficult because the systems and policies currently in place are not appropriate for admitting students with different academic backgrounds into the same study program for transdisciplinary research. However, the success of this doctoral college proves that this is how we must perform research in the future to overcome the existing silos of disciplines. The college has inspired certain changes that have been implemented in the Doctoral School of TU Wien and can serve as a model for future research projects. The introduction of the lecture "Responsible Research" for all doctoral students at TU Wien is one of the Trust Robots Doctoral College's most important accomplishments from our perspective. In this lecture, we consider ethical standards and the societal effects of innovation and science while preparing our students with morally sound design and trustworthy research techniques.

The doctoral college "Trust in Robots" is an unqualified success for TU Wien. The results of a four-year project at TU Wien are summarized in the twelve chapters of this book, which we are happy to release to the public.

Kurt Matyas (Vice Rector for Academic Affairs)
Johannes Fröhlich (Vice Rector for Research and Innovation)

Editorial

Robots are gradually becoming a part of our daily lives, populating our living and working spaces. We hope that robots will come to relieve us from chores and dangerous, dull, or dirty work. We believe that they can make our lives more comfortable, easier, and even more enjoyable by providing companionship and care. Hence, robots will change how we collaborate and assign tasks to human and machine agents and even—more fundamentally—how we live and perceive ourselves and our roles in society. Although we believe we have control over the machines we have built, this belief may fade as devices become more significant, autonomous, and influential. The independent actions of robots can be frightening. Thus, developing technology for people requires that we are—at all times—in control of the technology or that we can rely on the good intentions and safety of autonomous systems over which we have no control. Therefore, building trust in (autonomous) robot systems is necessary.

Trust has been an essential issue in automation and technology research since the 1980s. According to studies on interpersonal trust, trust as an attitude develops into reliance and so plays a crucial role in technology acceptance and appropriate use of automation. Furthermore, research indicates that the same social heuristics used in human–human interactions may apply to human–robot interactions (HRIs) because robots trigger similar social attributes as humans. Although previous research revealed disparities between trust in and reliance on technology and trust among people, this difference may become more blurred as robots increasingly mimic human interaction patterns and exhibit anthropomorphic appearance and behavior.

This problem is reflected in the title of this book, “Trust in robots—Trusting robots,” which carries different notions and unifies various research areas. While “Trust in robots” addresses the subject of how to develop technology that users are willing to rely on, “Trusting robots” focuses on the process of establishing a trusting relationship with robots, thereby extending previous research. This latter interpretation of trusting robots—although still to a great extent futuristic—poses the question of how to develop artificial intelligence and robotic technology that allows a robot to exhibit trusting skills when interacting with humans. It considers that humans may develop relationships with robots that go beyond technology acceptance and reliance. Thus, trust in this context does not only refer to the one-sided confidence of users toward robots but also to users’ need to be assured that robots incorporate notions of the meaning of objects and social norms, including biases, and have an understanding of scenes and situations to be capable of interacting with users socially. However, the mere possibility that we may develop bonding and emotional attachment to machines raises several ethical questions and concerns. Is it ethical to design devices that trigger trust and relationship building? Should robots simulate trustworthy behavior to start reciproca-

tion by their users? Does trust in robots increase the vulnerability of users? How can we increase transparency regarding the capabilities of robots to ensure that users understand what robots can and should do? Should robots mimic other human qualities—such as empathy or emotions—to enhance trust?

These questions and topics have been the core of the “Trust Robots” doctoral college at TU Wien. The main aim was to comprehensively analyze trust in the context of robotic technology from various perspectives. The book presents the results of the 4 year endeavor of doctoral students—from fall 2018 to fall 2022. Before summarizing their contributions, let us briefly discuss the critical scientific challenges in transdisciplinary research.

Scientific Challenges and Transdisciplinary Research

On the one hand, building trust in robot systems entails endowing robots with capabilities and skills to perceive and understand human communication and behavior (for example, through natural language processing, by recognizing facial expressions, voice, gestures, and emotions); to recognize and ideally predict human intentions; and to adequately respond to all of these stimuli. Furthermore, any robot reaction must guarantee users that they are safe at all times and that human rights are respected and ensured. On the other hand, humans must perceive robots as safe and reliable. Since it is impossible to foresee or enumerate all possible situations, autonomous (social) robots must respond securely to unexpected and unforeseen encounters. They must be able to learn and adapt, as they will be tasked with making independent decisions that go far beyond the preprogrammed security rules and algorithms. In such a context, robots are ascribed and will have (social) agency.

To address these research issues, researchers from different disciplines must collaborate to pool their expertise, methodologies, and knowledge. The envisioned assistance from robots to improve the quality of life and work can only be realized responsibly when the issues associated with this technology are considered appropriately. Consequently, there is a need to discuss and understand possible future scenarios from different perspectives: technological (i.e., implementing aspects of trust on robots), human (i.e.,., deployment of trustworthy robots in work and social contexts), and societal (i.e., legal, ethical, political, and sociocultural aspects).

Our research is based fundamentally on the sociomateriality paradigm, which holds that sociocultural processes and technology and its applications are inherently entangled and cannot be analyzed separately.

Furthermore, since industrial and social robots are intelligent, autonomous machines that lack moral capacity, scientists and developers must assume re-

sponsibility for ethically aligned design from the outset. Hence, ethical robotics begins with R&D rather than mitigating the adverse effects and harm caused by new technologies after they are introduced. Therefore, our research is guided by the principles of the responsible robotics paradigm and focuses on the ethical concerns associated with the incorporation of robots into society.

The faculty and students of the doctoral college are truly interdisciplinary: they have backgrounds in the philosophy of science, design science, labor science, economics, social science, psychology, computer science, mechanical engineering, and electrical engineering, and they have worked on 12 different topics arching from a principle design-perspective on sociotechnical systems over joint attention and motion planning to adaptive task sharing in human–robot collaboration and a general reflection of trustworthy robots in society. Figure 1 shows an overview of the transdisciplinary research at the doctoral college.

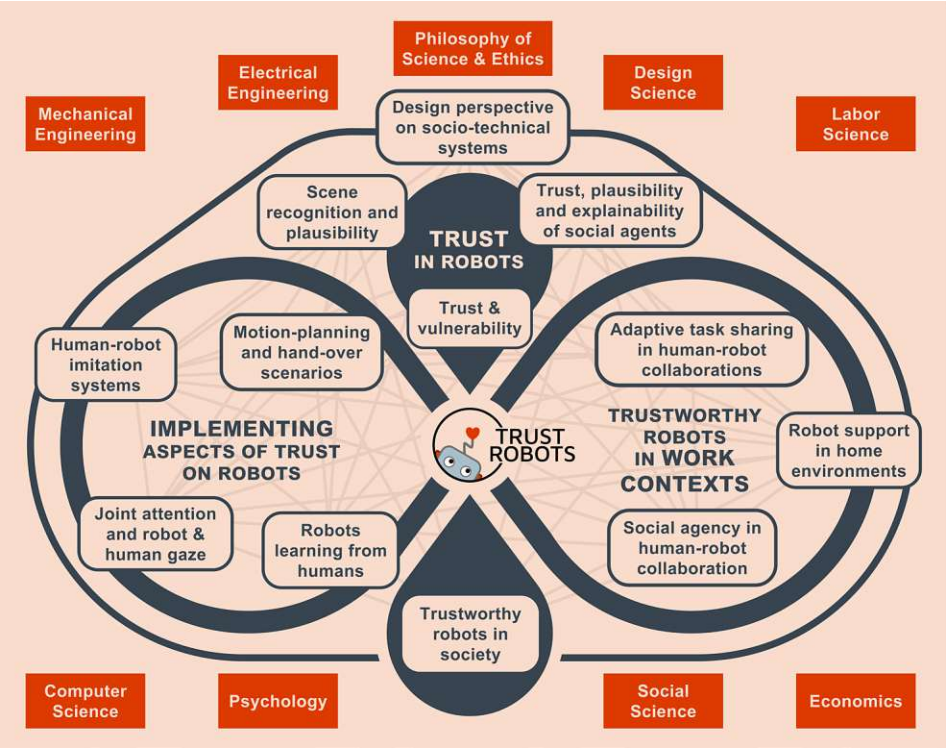


Figure 1 Transdisciplinary fundamental technical and applied research on implementing aspects of trust in robots

This work completed in the doctoral college is genuinely transdisciplinary. Students from different disciplines collaborated to develop implementations on robots, design experiments and demonstrations, analyze data, and draw conclusions from the findings for the field of HRI and their core disciplines. This

has led to the profound understanding that studying robots requires considering the entire sociotechnical system and context. This comprehensive perspective allows for designing meaningful and ethical robotic technology that will meet our expectations of making our lives easier and more enjoyable.

Summary of Results

The collection of articles in this book presents the highlights of the work on trustworthy robotics. We divide the summary into five sections: designing trustworthy robots, discussing trust and plausibility, implementing aspects of faith in robots, proposing that trustworthy robots must be viewed in the work context, and suggesting that trustworthy robots should be regarded in society.

Designing of Trustworthy Robots

In the first chapter of this book, Frijns & Schürer analyzed the contributions and importance of design work in the field of HRI research. They proposed that how interaction is conceptualized fundamentally impacts the design space and hence has to be considered in robotics research. Frijns et al. convincingly argued that the design space(s) for HRI must be extended beyond the individual aspects of humans and robots and encompass the sociotechnical system for which the robot is built. They make significant contributions to HRI through these design practice lenses.

Trust and Plausibility

The practical value of trust is founded on previous research findings that trust facilitates technology acceptance. Hannibal, Weiss & Purgathofer expanded on the perspective of “Trusting Robots” by providing a systematic identification of situational, robot-specific vulnerabilities in HRI. Hence, Hannibal, Weiss & Purgathofer shifted our focus to the contextual setting in which HRI occurs, challenging the prevalent negative association between interpersonal trust and vulnerability from both a theoretical—philosophical—and empirical perspective.

Based on the same fundamental idea of the relevance of context for HRI, Papagni & Koeszegi argued that for robots to be accepted within society, nonexpert users must find them valuable and trustworthy. They proposed to design robots that explain their decisions and actions to nonexpert users within the context of everyday interactions. Furthermore, they propose a model in which the plausibility of explanations resulting from contextual negotiations between the parties involved determines the understanding and supporting trust.

Bauer & Vincze applied this plausibility of explanations to the concrete scenario of scene interpretation, a core element of robots interacting in the world

and with people. It first presents the technical approach to creating an object hypothesis using learned methods and then employs a verification process to obtain relationships between objects in the scene. The work shows that such scene-level information should be used to estimate object poses. Their primary assumption is that all object hypotheses concerning their visual observation and the physical scene in which they reside must be plausible. These scene interpretations are then employed in reasoning strategies to explain to the user what the robot perceives during HRI.

Implementing Aspects of Trust in Robots

The following studies focused on how to implement these various aspects of trust in robots and trusting robots into technology.

Stoeva & Gelautz presented a framework for a human–robot imitation system and examined the system requirements imposed by different interactions for communicative, functional, artistic, or abstract movements. The analysis identifies open challenges for designing and developing human–robot imitation systems, such as the difficulty of observing and accurately replicating human motions and how to transfer human to robot motions given different embodiments (correspondence) and even measuring the deviations. The study also addresses ethical issues, such as keeping privacy, not deceiving interactants, and correctly employing the robot system as intended and agreed upon.

Following the interpretation of human gestures, the robot might contact the human, as in a hand-over scenario. Beck & Kugi investigated motion planning specifically for such trustworthy human–robot collaboration, emphasizing the significance of ensuring human safety and comfort during the interaction. Concerning comfort, the study emphasizes fluency (a high level of coordination between humans and robots, resulting in accurately timed, and efficient sequences of action), legibility (a measure of how well the robot conveys its intent), and human-like motion. The study introduces a receding horizon trajectory optimization approach to achieve such behavior, where the requirements for safety and comfort during the interaction are formulated in objective functions.

Another critical aspect of fluent interaction is for the robot to understand the intention of the human user and to build a mutual understanding of the subsequent actions. Koller, Weiss & Vincze studied this joint attention perspective using a robot and human gaze behavior during collaborative actions. The study reviews research on joint attention and the theory of mind as foundational elements for the success of collaborative tasks in human–human interaction. The authors employ the research approaches of roboticists to provide robots with a joint attention capability or at least the technically feasible equivalent. The idea is that mechanical gaze behavior, which humans can easily comprehend, will improve the inter-

action capability of a social robot. This is evaluated in an already established HRI joint action benchmark scenario of collaboratively building a tower out of different blocks.

Finally, it would be great if robots could continue to learn from humans in everyday life scenarios. To achieve this, Hirschmanner & Vincze proposed using a grounded language learning approach to connect words and references in social spaces, such as objects. The authors presented a Pepper robot-based incremental word learning system. Then, they introduce how to learn specific low-level activities through demonstrations. Furthermore, they present systems with an industrial robotic arm and a dexterous robotic hand as concrete examples. Additionally, they address the role of the teacher in the learning process, determining which human factors are essential to facilitate the learning process.

Trustworthy Robots in Work Contexts

As previously stated, developing trustworthy robots requires considering the system's context. Thus, we must also study the context in which robots are deployed. The imagination of the role of robots is often driven by technology and top–down ideological agendas, without regard for the practical realities of everyday life and work contexts. Schwaninger, Weiss & Fitzgerald explored bottom–up HRI research in the context of home environments and robot support for older adults. Furthermore, the study presents an overview of assistive technology for home environments, the building blocks for HRI research in these contexts, and the issues of elderly support and care.

Zafari & Koeszegi addressed questions regarding the extent to which robots are accepted in work settings, as well as the impact human–robot collaboration has on workers and their perceptions of their own and the robot's role, agency, and efficacy. They show how agency is ascribed to nonhuman entities and present two experiments that analyze this impact. Zafari et al. provided valuable recommendations for both the design of artificial agents and organizational strategies in terms of which social practices and changes in the working context must provide opportunities for a successful collaboration.

Schmiedbauer & Schlund addressed another essential aspect of successful human–robot collaboration: how to allocate tasks between humans and robots. Instead, of automating all that can be automated and leaving the rest to humans, they employed a human factors approach and focused on the needs and capabilities of workers and economic targets at the center of analysis. They designed, developed, demonstrated, and evaluated a model for adaptive task sharing between humans and cobots (collaborative robots) and showed avenues for further development based on their insights.

Trustworthy Robots in Society

Finally, DePagter provided a macrolevel analysis, i.e., an analysis of the process of building trust in robots on a societal level. They proposed a narrative approach and argued that robots are a prominent example of a technology that has caught many people's imagination of the future. The analysis of these future imaginaries of robots provides a deep understanding of how technology is perceived by the general public, what fears and hopes are associated with this technology, what roles are given to robots, and what challenges are associated with them. This narrative approach provides avenues for policymakers and developers to shape future imaginaries of robots.

Sabine T. Koeszegi, Markus Vincze

Table of Contents

List of Abbreviations

2D	Two-Dimensional
3D	Three-Dimensional
AAL	Active and Assisted Living
ADD	Average Distance of Model Points
AI	Artificial Intelligence
ANN	Artificial Neural Networks
ANT	Actor-Network Theory
AR	Average Recall
ATS	Adaptive Task Sharing
AUC	Area Under the Curve
BPMN	Business Process Model and Notation
CCVSD	Care Centered Value Sensitive Design
CHOMP	Covariant Hamiltonian Optimization for Motion Planning
CNN	Convolutional Neural Network
CPPS	Cyberphysical Production System
CSCW	Computer Supported Cooperative Work
CV	Computer Vision
DDP	Differential Dynamic Programming
DDPG	Deep Deterministic Policy Gradient
DOF	Degrees of Freedom
DTMC	Discrete-Time Markov Chains
EC	European Commission
EED	Eye-Direction Detector
EJA	Ensuring Joint Attention
EU	European Union
GDPR	General Data Protection Regulation
GJK	Gilbert-Johnson-Keerthi

GMM	Gaussian Mixture Model
GUI	Graphical User Interface
GuSTO	Guaranteed Sequential Trajectory Optimization
HCI	Human-Computer Interaction
HER	Hindsight Experience Replay
HF/E	Human Factors / Ergonomics
HHI	Human-Human Interaction
HRC	Human-robot Collaboration
HRI	Human-Robot Interaction
HSR	Human Support Robot
ICP	Iterative Closest Point
ICT	Information and Communication Technology
ID	Intentionality Detector
IJA	Initiating Joint Attention
IL	Imitation Learning
IRL	Inverse Reinforcement Learning
ISO/TS	International Standards Organization / Technische Spezifikation
LfD	Learning from Demonstrations
LIDAR	Light Detection and Ranging Sensor
MCTS	Monte Carlo Tree Search
ML	Machine Learning
MPC	Model Predictive Control
MDP	Markov Decision Process
MTM	Methods Time Measurement
MTM-AUS	Methods Time Measurement - Universelles Analysier-System
npmi	normalized pointwise-mutual information
OMPL	Open Motion Planning Library
PD	Participatory Design
PDDL	Planning Domain Definition Language
PPO	Proximal Policy Optimization

STOMP	Stochastic Trajectory Optimization for Motion Planning
STRIPS	Stanford Research Institute Problem Solver
STS	Sociotechnical Systems
SUS	System Usability Scale
SVM	Support Vector Machines
ToM	Theory of Mind
ToMM	Theory of Mind Mechanism
UCB	Upper Confidence Bound
UI	User Interface
UX	UsereXperience
VSD	Visual Surface Discrepancy
XAI	Explainable Artificial Intelligence

Designing of Trustworthy Robots

Design as a Practice in Human-Robot Interaction Research

Helena Anna Frijns, Oliver Schürer

Abstract

This chapter reflects on the scope, methods, knowledge contributions, and normative orientation of design for the research field Human-Robot Interaction (HRI). The design space of interactions between humans and robots is characterized as being influenced by the way interaction is understood. Underlying views of interactions merit consideration, as they influence the research questions, methods, and aims of HRI design research. It is argued that we need to understand the concept of the design space(s) for HRI as extending beyond individual aspects that can be varied in the design of interactions between humans and robots to encompass the socio-technical system that the robot is developed for. This chapter further characterizes the practice of HRI designers as comprising multiple overlapping activities, operating in a complex problem context in a design team with multiple sets of expertise from different disciplines, comparable to or functioning as transdisciplinary research. This chapter contains a discussion of knowledge contribution that can be achieved through design practice and concludes with reflections on the responsibility of designers.

Keywords

Human-Robot Interaction, Interaction Design, Design Research

1 Introduction

Human-Robot Interaction (HRI) is a multidisciplinary research field that integrates disciplines such as engineering, psychology, sociology, philosophy, and more[1]. Collaboration between different disciplines is necessary to achieve goals (such as developing robotic systems for human-aware navigation), but can be complicated as each discipline has a different jargon, uses different methods, and knows different practices and paradigms. Design is frequently described one of the disciplines of relevance in HRI. Lupetti et al. define designerly HRI as "*the body of work in HRI that has a strong orientation toward design (i.e., work developing novel robotic artifacts and/or engaging with design methodologies)*" [2021, p. 389]. They consider designerly HRI as a methodology or form of research, a "means for investigation" [Lupetti et al. 2021, p. 381] extending beyond individual robot designs or designed features. As it is necessary to collaborate across disciplines, this chapter aims to further clarify the role of (interaction) design in HRI and how it contributes (both in terms of knowledge and prototypes) to HRI research and the development of robotic systems.

This chapter is a position statement and literature review on the scope of the HRI design practice, activities that are part of design practice, the potential of design to contribute knowledge, and the normative orientation that design work

1 Key characteristics of disciplines include that they have a specific focus on certain phenomena, concepts, methods and theories, and that they subscribe to particular 'rules of the game' and specific disciplinary perspectives [Szostak et al. 2016, p. 10].

https://doi.org/10.34727/2022/isbn.978-3-85448-052-5_1

implies. Whereas design of robot appearances and behaviors for interactions between humans and robots has been a topic for several decades (see for instance the special issue on design for HRI [Holmquist and Forlizzi 2014] and work on social robot embodiment design and anthropomorphism [Blow et al. 2006; Hegel 2013; Deng et al. 2018]), and to an extent design methodology (for instance [Bartneck and Forlizzi 2004; Drury et al. 2004], and work on Value-Sensitive Design [Dignum et al. 2018; Van Wynsberghe 2016; Cheon and Su 2016]), recently there is an uptake of interest in reflecting on design methodology for HRI and how design research can contribute knowledge to the HRI community. This is exemplified by a series of recent papers and workshops on topics such as designerly HRI [Lupetti et al. 2021, 2020], integration of User eXperience (UX) design in a human-robot interaction design workflow [Prati et al. 2021], use of metaphors to inspire HRI design [Alves-Oliveira et al. 2021], combination of UX design and ethics in the design of social robot behavior [Fronemann et al. 2021], Research through Design (RtD) [Luria et al. 2021], exploratory prototyping for HRI [Zamfirescu-Pereira et al. 2021], and Design-Centered HRI and Governance [Weng et al. 2021]. Questions relevant to these workshops and papers include what an HRI design epistemology could be, evaluation of knowledge resulting from HRI design practices [Lupetti et al. 2020, 2021], and reflection on which RtD methods are relevant for HRI [Luria et al. 2021]. The recent interest in design methodology, design practice for HRI, and the necessity to collaborate across disciplines in HRI make the topics of design practices and design knowledge both timely and relevant.

This chapter references work in Human-Computer Interaction (HCI) and theory on design research that is relevant for HRI designers, as there is a certain maturity in those discussions that will be informative. This chapter seeks to answer several questions: why is design relevant for HRI? What can it be useful for? What do designers know or what can they do that can contribute to solving problems? Why is design positioned (perhaps uniquely positioned) to solve specific problems?

This chapter discusses the concept of design space(s), characterizations of the practice of designers, and characterizations of knowledge contributions that design can offer. It reflects on the ways of thinking about the activity of designing as part of an HRI research practice. Finally, the chapter argues that responsibility is inherent to the design practice as a result of one of the main aims of design, namely to change or impact people, societies, and the world.

2 The Design Space of Human-Robot Interactions - From the User Interface to the Socio-Technical System

2.1. The Concept of the Design Space

In this section, we discuss various levels at which we can consider design. What can designers affect? What is part of the "material" of a designer's practice? One concept we can start with is that of the design space. A design space can be described as the set of possible design alternatives or aspects of a design that can be altered. The term design space is frequently used to indicate that the design problem, object or system has various features that can be varied: design decisions have to be made regarding these features. The concept is commonly used in computational design, and it is gradually making its way into HRI as a way to describe a design problem.

Halskov and Lundqvist elucidate the concept of design spaces in a HCI context: *" (...) a design space may be represented in a number of ways, such as a Cartesian space, a network graph, or a conceptual space. The scope of a design space ranges from a class of technologies, over all accumulated knowledge during a specific design process, to the design space of a collection of designs, ideas, and sketches."* [2021, p. 3]. They note that the term design space can refer to the physical space where design activities take place. Thinking of a design problem in terms of its design space can also take the specific form of representing design aspects computationally, with requirements that need to be satisfied represented as objective functions that need to be optimized. Design space exploration refers to the idea that a large space in which designs are represented in a specific way can be traversed computationally [Woodbury and Burrow 2006]. Computational methods for exploring the design space can be useful for finding a solution that satisfies design objectives while exploring more of the design space. It assumes that the problem can be modeled as a combination of parameters to be adjusted to satisfy constraints [Chan et al. 2022]. A computational approach to the design space concept can be useful for restricting the problem scope and finding new design solutions within said restricted problem space. However, certain requirements or constraints are not easily (or at all) possible to represent as an equation or numerical condition/value that can be met. These operate at different levels, e.g. in interaction with one or multiple users, or only become apparent when the technology is introduced to society on a large scale.

The design space term can also refer to a metaphorical space containing possibilities and alternatives that are taken into consideration to satisfy design requirements [Halskov and Lundqvist 2021]. Botero et al. describe the design

space as "*the space of possibilities for realizing a design*" [2010, p. 1] and "*the space of potentials that the available circumstances afford for the emergence of new designs*" [2010, p. 3].

In the context of HRI, the design space term is often used in its conceptual or metaphorical sense. Deng et al. [2018]. describe the design space in terms of changes that can be made regarding the appearance, behavior, and structure of interactive technologies such as social robots. Baraka et al. [2019] propose a framework with seven dimensions for characterizing social robots, which they describe as forming a design space. The framework contains the following dimensions: appearance, social capabilities, purpose and application area, relational role, autonomy and intelligence, proximity, and temporal profile. The design spaces sketched by Deng et al. [2018] and Baraka et al. [2019] have a strong focus on the robot as a socially interactive device with a specific appearance and function. Other frameworks have been developed for describing HRI. Goodrich and Schultz [2007] describe the dimensions that HRI designers can affect (autonomy, information exchange, team structure, adaptation and learning, and task shape) with a focus on human-robot teamwork. In the HCI context, Forlizzi and Ford [2000]'s design framework of user-product interaction includes the user, the product, context of use, social and cultural factors. On the human side they include the factors emotions, values and prior experience; on the product side they include aesthetic qualities, form language, features, and usefulness.

2.2. Interaction Design, UI & UX

In discussing the main topic of design for HRI, the focus of the current chapter is on interaction design, that is, designing for interaction. Interaction design has been described as "*the shaping of digital materials — software, electronics, telecommunication, etc. — with a particular focus on the use of the resulting digital artifacts*" [Löwgren 2007, p. 1].

An interaction designer affects the appearance of a system, its behavior in response to stimuli, and the quality of interaction and User eXperience (UX) in a way that fits the context, with the aim to improve a current situation by changing existing systems and creating new systems [Smith 2006; Fallman 2008; Goodman et al. 2011]. A host of aspects can be considered; the control method, the usability of the user interface (UI), familiarity, timeliness and correctness of action execution by the system, clarity of communicative cues used by the system, how well the system recognizes human cues, which cues the system can recognize, information quality, the embodiment of the robot, aesthetic qualities, and so on. Interaction design can be considered at different levels, from the micro level of button clicks on a graphical user interface (GUI) to the macro level of societal

effects of technical systems. Although the scope of a design project may be restricted to, for instance, certain aspects of embodiment design, we would consider it of high importance that the setting that the robotic system is designed for, as well as broader ethical and societal implications, are taken into account during the design process.

First, the 'micro' level of human-technology interaction will be discussed here, by starting with the classical HCI design topics of UI and UX design before expanding the discussion to include a broader perspective on designing interactions, to argue that all these levels need to be considered in the design of technical systems such as robots.

To start off, we consider the UI. User interface design is highly relevant when discussing the topic of designing interactions with technical systems such as robots. UIs have been described as components or mechanisms that enable two-way human-machine communication, presentation of information, and human control of systems and processes to achieve specific tasks [International Organization for Standardization (ISO) 2010; Marvel et al. 2020]. The UI can also be described as all the means of input and output that offer humans interacting with the system the possibility to obtain information from a robot and affect the technical system across different modalities. This can include a GUI, motor sounds, gestures, sound alerts, and movement. Applying the concept of the user interface as familiar from other interactive technologies such as computers and smartphones is problematized in the case of HRI [Frijns et al. 2021]. Especially in the case of co-located robots, a human interacting with a robot will gain information from the robot via many other channels than just a GUI or other parts of the system that have been intentionally designed to convey information to an end user and allow an end user to act on the robotic system, as the embodiment of the robot and the way it moves and sounds (perhaps even smells and tastes) are informative and can be impacted.

Conceptually, restricting the UI to the input/output devices or mechanisms specific to the system renders the interaction rather flat, as interaction can never be just restricted to operations on the UI - the interaction is connected to the person, system, situation, and the world in addition to those specific input/output mechanisms employed in the UI. Consider for instance the concept of so-called *intuitive* use, a process that involves prior, partially automatic nonconscious knowledge (familiarity) [Naumann et al. 2007]. To achieve such a high level of ease of use, the design of the UI has to appeal to previous knowledge of the user, in other words, it should appeal and be connected to culture, prior experiences, motor memory, and so on. As soon as we talk about designing a UI, or consider a UI in interaction, we need to take the broader context into account, including one or

multiple users, other people (including other stakeholders), objects, and technologies.

Where the UI is conceptually restricted to the input/output aspects of an interaction and information exchange, the concept User eXperience (UX) is intended to cover the more experiential aspects, which still leaves some space for considering the whole experience of the system as well as unintended inputs and outputs, such as motor sounds, which are generally not intended to be part of the UI but do provide information to an end user. UX can be described as part of the design space of human-robot interactions, focused on the experience of interaction of an end user. Perceptions and understanding of and responses to (anticipated) use of the system, suitability to the context, and how the system serves human needs are seen as part of UX [International Organization for Standardization (ISO) 2010; Weiss et al. 2009]. Taking UX into consideration as part of the design space of social robots already accounts for more aspects than just looking at operations on a UI, but in focusing purely on the user and their experiences, it is clear that more aspects need to be considered when designing HRI systems - other effects and actors not considered in the UX concept.

2.3. Waves of HCI and Views of Interaction

The different ways of approaching the design space of interactions between humans and robots depend on the ways interaction itself is viewed. Harrison et al. [2007] discuss Kuhn's concept of the paradigm shift, and argue that similarly, HCI is characterized by paradigms that are dependent on the paradigm's metaphor for interaction. These metaphors influence the goals for the interaction, research questions that are asked, and the methods used. Consequently, research that is conducted with different foundational views of interaction subscribes to different epistemological bases.

In the HCI community, several shifts in focus and thinking have been identified and described as the *three waves of HCI*. Work that is conducted within (or across) such ways of thinking is informed by particular views of interaction. During the first wave, cognitive science and psychology were adopted as a way to inspire technology design, with a focus on information processing, human factors and model-driven thinking. The second wave entailed a shift from disembodied single-user interaction to collaborative communities working in a particular context, but still with a focus on users, exemplified by for instance the use of participatory design methods. During the third wave the focus shifted to design-oriented, more exploratory, critical, value-oriented technology development for daily life acknowledging the importance of such things as complexity, experience, meaning, and emotion [Bødker 2015; Fallman 2011; Harrison et al. 2007; Frauenberger 2019].

The metaphors reported by Harrison et al. [2007] as central to each wave are interaction as [hu]man-machine coupling, as information communication, and as phenomenologically situated, respectively. Where most authors distinguish three waves, Frauenberger [2019] proposes a fourth wave called *Entanglement HCI*. According to Frauenberger, HCI researchers/designers cannot "*design interaction*"; instead, they work on "*configuring material conditions*" [2019, p. 12].

Several theories, frameworks and accounts have attempted to describe what happens in the interactions between humans and technology, for instance the Product Ecology [Forlizzi 2008], Actor-Network Theory (ANT) [Law 1992], Activity Theory [Bertelsen and Bødker 2003], distributed cognition, and computational rationality [Oulasvirta et al. 2022]. *Interaction* can be understood or framed in different ways, as demonstrated by Hornbæk and Oulasvirta [2017] and by Frijns et al. [2021]. For instance, interaction has been conceptualized in the context of HCI as dialogue, transmission, tool use, optimal behavior, embodiment, experience, and control [Hornbæk and Oulasvirta 2017]. Naumann et al. [2007] describe interaction as information and energy exchange. Interaction and communication in HRI can be described, for example, as the sending of signals, as communicative action, as joint action or as a dynamic system, and main ways of framing interaction include interaction as control and as social interaction [Frijns et al. 2021]. Besides dyadic models of HRI, the attention on non-dyadic interaction is increasing [Schneiders et al. 2022].

Inherent to describing a communication process is the consideration where the communication is "located", or the question who is participating. What is social here, the relation between a human user and one or multiple robots, the relation of a user to the system's designers/developers, or social interactions that the robot enables between other agents? We can describe this as *sociality in the artifact, sociality through the artifact, or sociality with the artifact*. Conversely, we may describe *sociality as located across a network*, as in ANT.

Breazeal [2003] and Fong et al. [2003] distinguish several paradigms for social HRI that range from robots being *socially evocative* systems to robots being *socially intelligent*. Such paradigms are exemplary of a view of *sociality residing in the artifact* or as being a property of the robotic system or a human's relation with the robotic system: the artifact relates socially itself.

Another view of the communication process is that of the designer(s) of a system communicating with the end user, *sociality through the artifact*. For example, De Souza [2005] proposes semiotic engineering as a theory of HCI that construes computer systems as messages that are sent from the interactive system's designers to its users. The system functions as a deputy of the designer. It speaks for the designer, and this is described as a metacommunication process

- the designer's message is unpacked over the course of the user's repeated interactions with the system. De Souza states that computer systems encode a problem and a specific solution to that problem. Through exploration and negotiation of meaning with the system, the user is able to apply the designer's vision creatively to new problem situations.

Technologies such as robots can also be viewed as serving a mediating role; *sociality with the artifact*, where the artifact enables social relations between other actors. Such a role is described in the Domestic Robot Ecology [Sung et al. 2010]. See also the Product Ecology by Forlizzi [2008] and Raptis et al. [2014], who describe various ecology concepts that have been proposed within HCI, such as the information ecology, artefact ecology, and personal ecology. Van Wynsberghe and Li [2019] propose a reframing of the HRI model from dyadic interaction to a model of human-robot-system interaction (HRSI). A dyadic interaction model does not account for all the effects of introducing a robot in a care context, such as impacts on the healthcare system as a whole. In the model they propose, the bot is viewed as a mediator between the healthcare system and the patient. In this case, the bot is seen as closely connected to the company that developed it (for data collection, data processing, and upgrades).

Finally, *sociality can be located across a network*. Law [1992] characterizes ANT as a sociological approach that describes humans, machines, objects, organizations, society, and alike, as heterogeneous networks or the effects produced by heterogeneous networks. Actors are themselves networks (which is why actor and network are coupled in the name actor-network): "*(...) a machine is also a heterogeneous network - a set of roles played by technical materials but also by such human components as operators, users and repair persons*." [Law 1992, p. 384]. The concept of punctualizations describes the phenomenon that complex heterogeneous networks are masked by simple actions and that which causes the action, which comes to stand in for the complex network. This is applicable to a complex system such as a robotic system that comprises, for example, various devices and a human operator, but what is perceived is simply the robot performing actions. ANT scholars suggested that there is no distinction between the social and the material. Socio-materiality indicates that what is material constitutes the social, and the social constitutes the material [Leonardi 2012]. Yaneva [2009] discusses the application of ANT to design, arguing that design can be viewed as a connector that shapes social interactions. How something is designed is directly tied to the particular way in which it mediates social relationships; the way something is designed shapes the social in a particular way. Vallès-Peris and Domènech [2021] propose "Caring in the In-Between", an approach toward responsible technological development of robotics and AI technologies in the care

sector. The approach considers the robot as embedded in a network instead of as partaking in a dyadic HRI.

Alternatively, Verbeek [2008] describes technologies as not being neutral, and instead technologies serve a mediating role for human action, impacting human decision-making and configuring the conditions in which they can act and thus the conditions of their freedom. Verbeek distinguishes three forms of agency: human agency in the interaction with a technological artifact, the agency of the technology designer in shaping its mediating role, and the artifact's agency through the mediation.

These different views put the focus on developing different technologies. Contrast humanlike behavior for robots that relate to a human user in a humanlike social way to a view of social interaction as unpacking a designer's narratives in software (as in semiotic engineering), to a view of a robot impacting relationships within a family after its introduction to a household (e.g. Roomba [Sung et al. 2010]).

2.4. Design Spaces as Context-Specific

Harrison et al. [2007] state that the concept of design spaces fits the second paradigm or wave in HCI, as it suggests that there are aspects of design that can be varied without considering the context or how these aspects interrelate. A broader view of design spaces can be found with Botero et al. [2010], who write that the design space is not a pre-existing space, but instead, it is a co-constructed space formed by stakeholders, technologies, social processes. This moves the focus of the design activity away from the object, towards this broader context. This move towards including the context can and should also be made when discussing robotic systems, as "*(...) a system isn't complete without the people who use it*" [Smith 2006, p. xii] and the environment and situation it is embedded in. The concept of a design space should not be restricted to aspects that can be varied in isolation. Instead, it should be considered as situational and context-specific. Definitions of robots and robotic systems by the ISO focus on robots as programmable devices and associated sensors and other equipment [International Organization for Standardization (ISO) 2012]. However, in a design context, it makes sense to approach robots from a socio-technical systems perspective, as a system is designed for people. A broader approach can be found with the ISO definition of an interactive system, in which reference is made to hardware, software, associated services of the system, documentation, training, branding, and packaging [International Organization for Standardization (ISO) 2010]. One can go even further and include humans and their social worlds - and by extension, the natural environment.

Moving beyond a focus on individuals and their experiences, Frauenberger [2019] argues that the focus of design work should not be on designing better user experiences. Instead, designers should design for enabling "*meaningful relations*" within socio-material and socio-technical systems. Besides (or beyond) considering the impacts of technology on people, HCI should consider how humanity and its relationship to the world are reconfigured by technology design [Frauenberger 2019]. Johannessen and Perjons define a socio-technical system as "*a hybrid system that includes technical artefacts as well as humans and the laws, rules, and norms that govern their actions*" [2014, p. 12]. In order to design technical artifacts for socio-technical systems, a designer needs to recognize the knowledge present in such a socio-technical system and its individuals, practices and technologies. Though design as a discipline already moves beyond consideration of the technical artifact by itself, there is a need to consider effects on the situation and stakeholders involved, as well as larger societal implications. For social robotics, Šabanović similarly argues that it is important to ground robot design and the evaluation of robotic systems in "*real socio-technical ecologies inhabited by potential users*" [Šabanović 2010, p. 447], proposing the mutual shaping framework that acknowledges the mutual influence that robotics and society exert on each other.

To conclude, while the concept of the design space can be discussed in terms of aspects that can be varied, it is important to keep in mind that these aspects also have effects together, both on the interaction and at larger scales (e.g. organization, society). The design space of interactions between humans and robots can be approached in different ways, depending on the paradigmatic view of interaction that is subscribed to and where the interaction process is located *(sociality in, through, and with the artifact or across a network)*. Interaction can be approached in different ways (as control, or as social interaction) and at different scales or levels of impact, from clicking a button on a GUI to environmental effects from robotic e-waste. All these levels are more or less relevant depending on the focus and scope of the design problem. However, the existence of those levels should be kept in mind and the levels that are meant to be responded to should be specified. Interaction can be considered as actions using a UI, but this leaves many aspects of interaction unaccounted for. Although the concept of UX is broader, it still focuses on the experience of an individual user. Parallels can be drawn between a move from considering a design space as containing what can be observed locally in a specific interaction (e.g. in terms of actions on a UI) to a broader consideration of interaction as part of a socio-technical system, and the *waves of HCI*.

3 Activities, Methods, and Processes of Designers

3.1. Design Methods and Approaches in HRI

Several authors have studied and reflected on design practice in HRI (see also Section 1). Deng et al. [2018] note that three design disciplines are part of social robot design: interaction design, industrial design, and design of the animation of the robot. Baraka et al. [2019] distinguish three main design approaches in the context of social robot design, namely human-centered design, robot-centered design, and symbiotic approaches that take strengths and weaknesses of humans and robots into account to design for symbiosis. Alves-Oliveira et al. [2022] identify three types of design processes for social robot design. A linear process includes sequential steps, for example, hardware exploration followed by interaction design experiments, implementing expressive movement, interaction design, and then resolving conflicts in the design. An iterative robot development process involves continuous improvement of the system's design based on user and team feedback. Data-point-driven processes take insights, background knowledge, and experiences into account.

Design methods used in HRI listed by Lupetti et al. [2021] include animation studies, 3D modeling, sketching, brainstorming, and human-centered design methods such as interviews, questionnaires, participatory design methods, focus groups, observations, personas, and critical design. User involvement is important; Alves-Oliveira et al. [2022] write that if user needs are not met and designs are not sufficiently validated through user involvement, this runs the risk of applying stereotypes in the robot's design and experiencing pushback from end users and other stakeholders as a result. A process that involves users at different stages in the workflow can lead to a more holistic understanding. Such a process can involve multiple different methods, such as surveying, interviewing or observing target users.

3.2. Characterizing Design Research

Design research practice can be conceptualized in different ways. It can involve activities ranging from the design of specific instances and engaging in a design process, to the development of methods and generalization of knowledge derived from the design practice into theory in some form, while being informed by a design stance.

Design practice can be characterized as comprising several overlapping activities. Different conceptual levels on which designers operate can be discussed.

Fallman [2008] proposes a model for interaction design that depicts interaction design research as a triangle with *design practice*, *design studies* and *design exploration* at its corners. Interaction design activity is made up of combinations of activities from all three areas. Fallman describes *design practice* as *practicing design*, that is, developing products and prototypes in a design team informed by a specific design research question. *Design exploration* on the other hand, is directed toward searching for alternatives, criticizing the state of things, and taking aesthetics into account in interaction design research, which links the activity to practices in contemporary art. The aim of *design studies* is to develop a discourse or body of knowledge around design research and its results, aiming to generalize and understand [Fallman 2008]. The remainder of this section discusses literature on characterizing design practice in a way that corresponds to the set of overlapping activities discussed by Fallman, noting that many activities fit multiple domains.

3.2.1. Design Exploration

In contrast to design work that aims to meet certain functional, idealistic or market demands, design work can also be applied to ask questions rather than answer them. Designers can propose counternarratives, which may be one of the powerful things about design. Speculative design is not bound to market demands or aiming to serve a specific function besides the encouragement of societal debate. Critical design uses speculative design to critically question the status quo (e.g. preconceptions) regarding, for instance, the role of technologies such as robots in our life [Auger 2014]. This is one of the advantages that critical and speculative design offers; it enables stepping outside existing narratives and critically questioning them, and can be used to propose new narratives.

3.2.2. Design Studies

Zimmerman et al. [2010] characterize design theory as either *theory on design* (knowledge of design as an activity) or *theory for design* (knowledge developed to improve the design practice), whereas Research through Design (RtD) is "*a research approach that employs methods and processes from design practice as a legitimate method of inquiry*" [Zimmerman et al. 2010, p. 310]. Interest in RtD has increased as the focus has shifted in HCI from improving usability to designing for wicked problem situations.

3.2.3. Design Practice

In the HCI, design research and design science communities, multiple characterizations of practices of designers can be found. Different types of activities can be part of a designer's practice, and different ways of conceptualizing design work and its aims exist. Johansson-Sköldberg et al. [2013] describe different discourses on design, contrasting designerly thinking as found in the academic literature to the design thinking discourse within managerial discourse. They write that design thinking, in contrast to designerly thinking, equates creativity to the design practice (although there is more to the design practice), and that design thinking is viewed as a toolbox in a way that lacks context. Johansson-Sköldberg et al. [2013] distinguish five "sub-discourses" in the academic literature for designerly thinking and design, namely as "*creation of artefacts*" (Simon), as a "*reflexive practice*" (Schön), as a "*problem-solving activity*" for wicked problems (Buchanan, Rittel and Weber), as a "*way of reasoning/making sense of things*" (Lawson, Cross) and as "*creation of meaning*" (Krippendorff). These discourses have different epistemological origins [Johansson-Sköldberg et al. 2013, p. 124].

A dominant perspective is that of the problem-solving perspective on design, Johannesson and Perjons [2014] write that design research (and specifically design science) solves practical problems through the development of artifacts, that is, a system, method, model or otherwise that is intentionally developed towards an end. They write that many such problems are so-called "*wicked problems*". Rittel and Webber [1973] introduce the term wicked problems in relation to planning theory. For planning tasks it should be considered what would be the right thing rather than the most efficient thing to do. Planners encounter situations involving societal problems that are ill-defined, without a clear goal for the solution and unclarity if a solution that is found will actually solve the problem. Similarly, wicked problems in design thinking are characterized by Buchanan [1992] as problems that are ill-formulated, in an environment with multiple stakeholders, contradicting information and values, in which intervention can have unpredictable results. Dynamically changing requirements and conflicting, fragmentary knowledge can make such problems difficult to solve [Johannesson and Perjons 2014]. Buchanan writes that design as a discipline defies definition and that design does not have a specific subject matter, and rather, designers need to respond and relate to problems in the given circumstances, taking into account the views of stakeholders. The easily-forgotten process of making the product concrete in the wicked problem context is part of the domain of design, and the design process cannot be reduced to its final product alone [Buchanan 1992]. In Šabanović [2010]'s mutual shaping framework, social robot design is put forth as a wicked problem. Social robots are intended for applications in society, a problem context with increased uncertainty and complexity, which requires the design

to be more adaptable and requires more ethical consideration. Šabanović [2010] argues that new methods are required for social robot design that incorporate social and technical facets.

Several authors discuss complexity as part of the design practice [Stolterman 2008; Goodman et al. 2011], in line with the "wicked problem" narrative. Goodman et al. [2011] describe interaction design as a complex discipline involving different activities and types of knowledge and skills such as empathy with end users, technology knowledge, and capability of judging aesthetics. They describe a specific type of knowledge in the design discipline that rests on interpretation and reflective practice, with inherent ambiguity. The design practice is context-specific; from this context complexity arises and is experienced by the designer [Goodman et al. 2011]. Stolterman [2008] contrasts complexity in design to complexity in science and argues that these forms of complexity should be understood as different. Design complexity (or richness) arises from the designer's subjective experience in response to information, requirements, and possibilities in the situation that is to be designed for. While in scientific practice it is possible to reduce problem complexity by reducing the scope of the problem, for instance by only looking at very specific aspects of it, design practice needs to approach a situation holistically, which means that design complexity cannot be reduced in a similar way [Stolterman 2008].

Parallels can be drawn between ways of working in a design research team that aims to gather knowledge and develop solutions to the practical problems of a particular community and transdisciplinary research projects. In meeting the demands of a complex problem situation, both involve drawing on the knowledge of several academic fields and of stakeholders outside academia. With transdisciplinarity, the aim is to "*provide contextualized answers to complex questions*" [Szostak et al. 2016, p. 7], often by working in teams with several academic disciplines as well as non-academic stakeholders. In contrast, multidisciplinarity (or pluridisciplinarity, polydisciplinarity) involves the juxtaposition of several separate disciplines in terms of their methods and knowledge, without integration of those perspectives or developing a shared understanding [Szostak et al. 2016]. Interdisciplinarity has been defined as "*communication and collaboration across academic disciplines*" [Jacobs and Frickel 2009, p. 44]. With interdisciplinarity, the aim is to answer questions shared by several disciplines, integrating knowledge, theories and methods from these disciplines to develop a better understanding. This requires integration in an interdisciplinary research team; the team members should develop understanding of the others' perspectives [Szostak et al. 2016]. Reflecting back on the HRI context, HRI design research can be conducted in a multi-, inter-, or transdisciplinary fashion. Baraka et al. [2019] note that social robot design employs methods and approaches from the research

fields HCI, computer science, engineering and human factors. In an interview study with roboticists who worked at companies that manufactured social robots, Alves-Oliveira et al. [2022] describe that their interviewees all reported being part of interdisciplinary teams, including such disciplines as mechanical and electrical engineering, computer science, psychology, and the arts. Šabanović et al. [2007] write that social robots can function as "boundary objects" in the collaboration across disciplines, providing a common focus while also functioning as relevant research objects in individual disciplines. Blackwell [2015] argues that instead of thinking about HCI as a discipline, one might also frame the field as an *inter-discipline* or trading zone in which researchers work from an interdisciplinary standpoint, negotiating between and collaborating with different disciplines, instead of trying to consolidate it as a discipline by itself, spurring innovation rather than establishment of a body of knowledge. Blackwell describes HCI as practice-based, requiring collaboration and reflection. This can also be argued for design work in the field of HRI.

To summarize, design research can include activities such as theory development, design exploration and developing prototypes and systems in a design practice. Different perspectives on design practice exist, among which a problem-solving perspective is dominant. When design research is conducted with the aim to solve practical problems in a real-world context in a design team that draws on several sets of expertise from different disciplines, design practice can be characterized as operating in a transdisciplinary context to solve wicked problems. However, other characterizations of design work are possible, depending on the activities that are conducted and by whom the work is conducted.

4 Design Knowledge: From Ultimate Particulars to Global Knowledge Production

4.1. Ultimate Particulars vs. Global Knowledge Production

There is a tension between local, context-specific results from design work and the aim to derive knowledge from these results that generalizes across other situations or problem contexts. While produced artifacts can be studied as part of sciences, reasons Buchanan [1992], this is different from what happens in the design context; the easily forgotten process of making the product concrete in the "wicked problem" context is part of the domain of design, and the design process cannot be reduced to its final product alone. Design contributes localized, context-specific results, in the words of Buchanan: "...design is fundamentally concerned with the particular, *and there is no science of the particular*" [Buchan-

an 1992, p. 17]. Stolterman [2008] writes that design activity is aimed at creating (to enable) ultimate particulars, that is, each specific situation (system, organization, people, context) will result in a different outcome when a designed artifact is introduced, and the designer should consider the specifics of a particular use context, even if the designed artifact is the same. This is the direct consequence of designing for a specific socio-technical system as sketched before. Stolterman contrasts this with the aim of science, which is to "*formulate universal knowledge that explains the complexities of reality on a level removed from specifics and particulars*" [Stolterman 2008, p. 58] - which Stolterman notes is a crude description of scientific aims, nevertheless, this still serves to illustrate the contrast.

Design science, in contrast to localized design practices, is a field of research in which knowledge production through design is recognized as contributing to a global practice. Johannesson and Perjons [2014] write that results from design activities are at times relevant for a local practice only, whereas design science aims to produce results for a global practice (which effectively comprises multiple local practices and the research domain). They argue that for design results to become relevant to design science, research methods used must be rigorous, the resulting knowledge should relate to existing knowledge, and should be fed back into the community of researchers and practitioners whom this knowledge is relevant for. Stolterman [2008] warns against a design science approach to interactive system design as this may risk using methods that are not appropriate for design practice.

4.2. Intermediate-Level Knowledge

The concept of intermediate-level has been proposed to bridge/unify a field that has both local and global knowledge contributions, and everything in between. Lupetti et al. [2021] argue that current design work in HRI is usually restricted to design instances, but in order to build on findings in design research, researchers need to move beyond production of individual instances. They discuss design knowledge as resulting from a reflective practice (Research through Design); knowledge is produced by reflecting on the activity of designing or reflecting on resultant artifacts. They discuss the concept of intermediate-level knowledge, which occupies a territory between general theories and specific instances. They argue that this concept could be informative toward the development of a *HRI design epistemology*. The concept of intermediate-level knowledge is also part of interaction design and HCI discourse Höök and Löwgren [2012]. To count as academic knowledge contributions, contributions proposed as intermediate-level knowledge should fulfill the academic quality criteria of being contestable (contribution is not already generally accepted and can be questioned, which implies a

certain novelty to the contribution), defensible (rigorously argued) and substantive (relevant and worth the time investment) [Höök and Löwgren 2012]. Examples of intermediate-level knowledge include design guidelines, design methods, design patterns, and strong concepts. A criticism of Research through Design, as identified in an interview study, is that knowledge development was only implicitly part of the process or took place after project completion, and poor documentation of RtD processes [Zimmerman et al. 2010]. Additionally, RtD was critiqued on grounds of the existence of a romanticized view of the design process and the "genius designer" by practitioners and researchers engaged in RtD. Such a view may hinder knowledge development that is "*systematic, rigorous and relevant*" [Zimmerman et al. 2010, p. 316]. Lupetti et al. [2021] argue that design knowledge could be represented and built upon better in HRI if researchers would clearly document and articulate motivations regarding engagements with design activities.

Frauenberger [2019] criticizes the concept of intermediate-level knowledge, as it postulates the existence of a spectrum ranging from universal theories to individual design instances, that is, from positivism to social constructivism, without a shared epistemological basis. Frauenberger further criticizes the concept of intermediary knowledge for implying a loss of contextualized knowledge while not being sufficiently well-formulated to serve as theory: by looking for patterns of successful designs that can inform future designs, the context-specificity of what made it successful in the original configuration is lost. Besides, Frauenberger argues that there may be value not in how the pattern was similar in different contexts and thereby abstracting, generalizing, and reducing this design situation, but rather in how "enactments" were different. Intermediate-level knowledge may risk criticisms of lack of rigor or disregarding context [Frauenberger 2019]. Zamfirescu-Pereira et al. [2021] argue regarding generalization of design findings from exploratory prototyping for human-robot interaction design that the understanding of similarities and differences in application contexts is more relevant than the replicability of results.

A different spectrum of theory in design research can be found with Zimmerman et al. [2010], who view research results through design as contributing to theory through exploration. Zimmerman et al. [2010] argue that RtD has contributed knowledge in the form of "nascent theory", which can be placed at the start of a spectrum of knowledge development, where the spectrum ranges from nascent theory (resulting from exploratory work) to mature theory. Forlizzi [2008] describes the product ecology framework as a form of nascent theory, for instance. This framing suggests a "spectrum" from exploratory to more substantive theory contributions. This differs from the concept of intermediate-level knowledge as described before.

The concept of intermediate-level knowledge can be connected to a design stance and normative aspects of design. For example, design principles indicate values when they argue for such things as transparent communication to end users. Such principles inform an attitude to design and express a certain worldview: a particular reading (that may change in the future) of what design should do and what design artifacts represent and mean. There is a risk that intermediate-level knowledge in the form of design guidelines, for instance, can be interpreted as prescriptive, but the designer is also responsible in considering if and how such guidelines apply to a particular context. Besides its use to inform specific designs, it can serve as way to document a particular design stance/normative orientation, which can be useful for learning (developing a design stance oneself) as well as studying design research. Regarding the concept of the design stance, Buchanan [1992] considers two levels on which designers work: on a general level and on the level of a *quasi-subject matter*. The *quasi-subject matter* is part of the problem and situation at hand, and consists of a set of issues that is not exactly defined in the (wicked) problem context. The designer responds to the quasi-subject matter with a specific product, thereby making the quasi-subject matter concrete. The general level is explicitly described as not being constitutive of any kind of science, rather, it informs a kind of design stance or a general view of designed artifacts, the methods and scope of the design practice.

5 Design Research as a Normative Activity

- "*In contrast to empirical research,design research is not content to just describe, explain, and predict. It also wants to change the world, to improve it, and to create new worlds. Design research does this by developing artefacts that can help people fulfil their needs, overcome their problems, and grasp new opportunities.*" [Johannesson and Perjons 2014, p. 1]
- "*Everyone designs who devises courses of action aimed at changing existing situations into preferred ones*" [Simon 2008, p. 111]
- "*In essence, design is about understanding the current state and then designing an improved future state*" Holmquist and Forlizzi [2014, p. 1]

To summarize, a strong narrative regarding design research practice is that design concerns itself with building future situations - identifying needs/problems in a current situation and developing systems and artifacts that alter the situation, with the aim to improve it. In a problem-solving view of a design practice, the aim to improve an existing situation is a value judgement on what a preferred condition would be. This improved future state that design research is said to strive for could entail an improved user experience, better living conditions, or empowerment of users, though we may also go beyond the idea of "serving

user needs"; Frauenberger [2019] describes technology creation as a process in which humanity redefines itself. The aim of design is not "universal knowledge production" as a project in and for itself, as in science, abstracting reality while guaranteeing reproducibility and objectivity (see Stolterman [2008]'s description), from an observational standpoint. As established previously, design work is instead context-specific and calls on the subjective experience of the design team involved, as well as on others' subjective experience (e.g., that of stakeholders). Bartneck et al. write that designers (and engineers) aim to transform reality rather than understand it. Bartneck et al. consider the latter to be the aim of science [Bartneck 2020].

Transforming reality implies an *intentional stance*; designers have aims when designing artifacts and systems, such as supporting people in their work [Johannesson and Perjons 2014][2]. Buchanan writes: "*The history of design is not merely a history of objects. It is a history of the changing views of subject matter held by designers and the concrete objects conceived, planned, and produced as expressions of those views*." [Buchanan 1992, p. 19]. Technology developers have purposes for the work they do, whether such aims are explicitly stated, for instance, building efficient systems, or more implicit. Cheon and Su [2016] investigate narratives that indicated values in interviews with 27 roboticists. One of the motivations of roboticists they identified was to research (features of) humans such as human intelligence and language by developing humanoid robots. Šabanović argues that a designer's cultural assumptions impact robot design and identifies a technocentric mindset in which robots are viewed as "*technological fixes*" [2010, p. 439] (see also process dogma and the other oblique constraints for technology design identified by Auger et al. [2017]). Note that designers can also find themselves within an environment that produces a certain normative orientation. Rather than seeing robots as a technological fix, we should acknowledge that design comes with additional consequences. Technology opens up specific possibilities for action, potentially closing others.

Technologies mediate the way they are used; human action is directed, shaped, impacted by technology use. Verbeek [2008] posits that technological artifacts have a form of material morality. Technologies are not neutral; instead, they are "*active mediators that help shape the relation between people and reality. This mediation has two directions: one pragmatic, concerning action, and the other hermeneutic, concerning interpretation*" [Verbeek 2008, p. 94]. First,

2 In discussions of the three waves of HCI, questions have frequently been asked regarding what "good" means in relation to the third paradigm and what should be strived for in technology design. Fallman asks "what constitutes a good user experience" [Fallman 2011, p. 1053] and proposes taking philosophy of technology (especially Borgmann and Ihde) as a starting point to consider questions regarding what the vision of "good" may mean for third wave HCI. Similarly, Harrison et al. [2007] asks "what it means for a system to be 'good' in a particular context" [Harrison et al. 2007, p. 6].

this means that technologies influence and shape human action. Second, they bring awareness in the sense of offering the possibility for humans to interpret a given situation in a different way, and enable different choices than would be the case without said technologies (e.g. Verbeek gives the example of conducting an ultrasound and the possibilities for choice and action this opens up). Although the action of the artifact is not deliberate, it gives direction to human action. "*Technological mediation, therefore, can be seen as a specific, material form of intentionality.*" [Verbeek 2008, p. 95] What is noted is that the intentionality of the technological artifact cannot exist in isolation; rather, it arises from the combination of technological mediation with human decision-making (hybrid intentionality). Technological artifacts thus represent a kind of constitutive force for human action, implying that technological artifacts implicitly direct human action (thereby having a form of material intentionality) as well as configure (some of the) conditions for human freedom. Because technology configures material conditions and impacts people's decision-making and freedom, "*technology design is inherently a moral activity*" [Verbeek 2008, p. 99] and designers should concern themselves with the future roles of the technologies they are developing - even though it is difficult to predict how technologies will mediate human actions in different contexts.

The intentional stance of designers (and that of engineers, too) brings responsibility. Stolterman writes that "*research aimed at changing and improving "reality" always takes on responsibility in relation to whom or what it serves*" [Stolterman 2008, p. 63]. This responsibility is acknowledged in e.g. Value-Sensitive Design, which positions alignment with specific values to the forefront in a design process. For instance, the aim of Care Centered Value Sensitive Design (CCVSD) [Van Wynsberghe 2016] is to incorporate care ethics into care robot design. Fronemann et al. [2021] argue that for social robots, risks of loss of control and privacy should be investigated and argue that design solutions that address these risks can be found by combining UX design and ethics. The point remains that apart from integrating ethics/values into the design process, the aim of design work should be critically reflected upon.

As sketched in Section 2, it is important to consider designing for the socio-technical system. However, this discussion can be taken even further. Going beyond socio-technical systems, toward the socio-material conditions mentioned by Frauenberger, it is also necessary to give consideration to other biological species and the natural environment (as argued, for instance, in relation to AI ethics [Owe and Baum 2021]). Such a proposal towards seeing technology, society, and nature as entangled can also be found in *critical making*, which acknowledges the fundamental interconnection between nature and culture. As a society, we face social and environmental problems that need to be addressed in a way that acknowledges the hybridity of nature and culture, community values and Global

North-Global South relations. "*The stakes are (...) high - nothing less than the fate of our planet (...)*" [Ratto 2016, p. 28].

6 Conclusion

To revisit the line of argument followed in this chapter, it was argued that the socio-technical system that a robotic system is embedded in needs to be considered as part of the design space of interactions between humans and robots. The concept of interaction that is subscribed to merits consideration, as this informs the research questions that are asked, methods used, and solutions that will be proposed. Taking a view of design work as solving wicked problems, HRI designers operate in complex problem contexts, often requiring collaboration across academic and practical disciplines, in order to design/configure conditions for the socio-technical system that is the HRI design space. However, other approaches to the design practice are possible, for instance, design practice as reflection on or criticism of current situations.

We cannot conclude what "the design practice" "is", as it comprises many different activities, aims, and contexts, at different levels of detail. It is open-ended, transforming with the possibilities and demands of a specific situation and insights and design stance of designers who respond to this situation. From Fallman [2008]'s conceptualization of interaction design work and the complexity of inter- and transdisciplinary design work, we conclude that designers employ a variety of methods and (can and should) use multiple lenses within their "discipline". The different perspectives offered through a critical design approach, producing specific design instances in context, the implicit design stance that design professionals develop over the years, and design theory development can inform each other and function in complementary ways.

A tension exists between the "localized" knowledge contributions that design practices produce compared to global knowledge production in design science or design research. The concept of intermediary knowledge has been proposed by other authors to bridge those local and global results, but such a concept can be criticized if it depicts knowledge contributions as lying on a spectrum from specific to general knowledge contributions. However, what can be acknowledged is that there can be value in such knowledge contributions as documenting a particular design stance or interpretation of design instances.

Finally, it was argued that a designer's intentional stance is inherent to design work. Typically, the aim is to transform reality to a more desirable state (with what qualifies as desirable depending on those involved in the design process, for instance end users in participatory design), but other aims can include criticiz-

ing the current state (e.g., in critical design) or imagining a different state (e.g., speculative design). When HRI research is applied in practice, this makes the social responsibility on the part of HRI designers apparent. Designers also find themselves within an environment (e.g. institutions such as universities, corporate environments, academic discourse) that produces a certain normative orientation and introduces constraints. It remains important to reflect on one's social responsibility and how our institutions and discourses impact and reinforce normative orientations in relation to this responsibility.

Bibliography

Patrícia Alves-Oliveira, Maria Luce Lupetti, Michal Luria, Diana Löffler, Mafalda Gamboa, Lea Albaugh, Waki Kamino, Anastasia K. Ostrowski, David Puljiz, Pedro Reynolds-Cuéllar, Marcus Scheunemann, Michael Suguitan, and Dan Lockton. 2021. Collection of Metaphors for Human-Robot Interaction. In *Designing Interactive Systems Conference 2021* (Virtual Event USA). ACM, 1366–1379. https://doi.org/10.1145/3461778.3462060

Patrícia Alves-Oliveira, Alaina Orr, Elin A. Björling, and Maya Cakmak. 2022. Connecting the Dots of Social Robot Design From Interviews With Robot Creators. *Frontiers in Robotics and AI* 9 (2022), 1–15. https://doi.org/10.3389/frobt.2022.720799

James Auger. 2014. Living With Robots: A Speculative Design Approach. *Journal of Human-Robot Interaction* 3, 1 (2014), 20. https://doi.org/10.5898/JHRI.3.1.Auger

James Auger, Julian Hanna, and Enrique Encinas. 2017. Reconstrained Design. In *Nordes 2017, Design + Power* (Oslo, Norway). 8.

Kim Baraka, Patrícia Alves-Oliveira, and Tiago Ribeiro. 2019. An extended framework for characterizing social robots. *arXiv*:1907.09873 [cs] (2019), 1–4. http://arxiv.org/abs/1907.09873

Christoph Bartneck. 2020. Design. In *Human-robot interaction: an introduction*. Cambridge University Press,41–68.

Christoph Bartneck and Jodi Forlizzi. 2004. A designcentred framework for social human-robot interaction. In RO-MAN 2004. 1*3th IEEE International Workshop on Robot and Human Interactive Communication (IEEE Catalog No.04TH8759)* (Kurashiki, Okayama, Japan). IEEE, 591–594. https://doi.org/10.1109/ROMAN.2004.1374827

Olav W Bertelsen and Susanne Bødker. 2003. Activity Theory. In *HCI models, theories, and frameworks: Toward a multidisciplinary science*. 291–324.

Alan F Blackwell. 2015. HCI as an Inter-Discipline. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems* (Seoul Republic of Korea). ACM, 503–516. https://doi.org/10.1145/2702613.2732505

Mike Blow, Kerstin Dautenhahn, Andrew Appleby, Chrystopher Nehaniv, and David Lee. 2006. Perception of Robot Smiles and Dimensions for Human-Robot Interaction Design. In *ROMAN 2006 - The 15th IEEE International Symposium on Robot and Human Interactive Communication* (Univ. of Hertfordshire, Hatfield, UK). IEEE, 469–474. https://doi.org/10.1109/ROMAN.2006.314372

Andrea Botero, Kari-Hans Kommonen, and Sanna Marttila. 2010. Expanding Design Space: Design-In-Use Activities and Strategies. In *Design and Complexity - DRS International Conference 2010* (Montreal, Canada). 13.

Cynthia Breazeal. 2003. Toward sociable robots. *Robotics and Autonomous Systems* 42 (2003), 167–175. http://web.cecs.pdx.edu/~mperkows/CLASS_ROBOTICS/FEBR26-2004/Humanoids/sociable-robots-Breazeal-RAS03.pdf

Richard Buchanan. 1992. Wicked Problems in Design Thinking. *Design Issues* 8, 2 (1992), 5–21.

Susanne Bødker. 2015. Third-wave HCI, 10 years later—participation and sharing. *Interactions* 22, 5 (2015), 24–31. https://doi.org/10.1145/2804405

Liwei Chan, Yi-Chi Liao, George B Mo, John J Dudley, Chun-Lien Cheng, Per Ola Kristensson, and Antti Oulasvirta. 2022. Investigating Positive and Negative Qualities of Human-in-the-Loop Optimization for Designing Interaction Techniques. In *CHI Conference on Human Factors in Computing Systems* (New Orleans LA USA). ACM, 1–14. https://doi.org/10.1145/3491102.3501850

EunJeong Cheon and Norman Makoto Su. 2016. Integrating roboticist values into a Value Sensitive Design framework for humanoid robots. In 2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI). 375–382. https://doi.org/10.1109/HRI.2016.7451775

Clarisse Sieckenius De Souza. 2005. Semiotic engineering: bringing designers and users together at interaction time. *Interacting with Computers* 17, 3 (2005), 317–341. https://doi.org/10.1016/ j.intcom.2005.01.007

Eric C Deng, Bilge Mutlu, and Maja J Matarić. 2018. Formalizing the Design Space and Product Development Cycle for Socially Interactive Robots. In *Workshop on Social Robots in the Wild at the 2018 ACM Conference on Human-Robot Interaction (HRI)*. 6.

Virginia Dignum, Frank Dignum, Javier Vázquez-Salceda, Aurélie Clodic, Manuel Gentile, Samuel Mascarenhas, and Agnese Augello. 2018. Design for Values for Social Robot Architectures. *Envisioning Robots in Society - Power, Politics, and Public Space* (2018), 12.

Jill L Drury, Dan Hestand, Holly A Yanco, and Jean Scholtz. 2004. Design guidelines for improved human-robot interaction. In *Extended abstracts of the 2004 conference on Human factors and computing systems - CHI '04* (Vienna, Austria). ACM Press, 1540. https://doi.org/10.1145/ 985921.986116

Daniel Fallman. 2008. The Interaction Design Research Triangle of Design Practice, Design Studies, and Design Exploration. *Design Issues* 24, 3 (2008), 4–18. https://doi.org/10.1162/desi. 2008.24.3.4

Daniel Fallman. 2011. The new good: exploring the potential of philosophy of technology to contribute to human-computer interaction. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Vancouver BC Canada). ACM, 1051–1060. https://doi.org/10.1145/ 1978942.1979099

Terrence Fong, Illah Nourbakhsh, and Kerstin Dautenhahn. 2003. A survey of socially interactive robots. *Robotics and Autonomous Systems* 42, 3 (2003), 143–166. https://doi.org/10.1016/ S0921-8890(02)00372-X

Jodi Forlizzi. 2008. The Product Ecology: Understanding Social Product Use and Supporting Design Culture. *International Journal of Design* 2, 1 (2008), 11–20.

Jodi Forlizzi and Shannon Ford. 2000. The Building Blocks of Experience: An Early Framework for Interaction Designers. In *DIS'00* (Brooklyn, NY). 419–423.

Christopher Frauenberger. 2019. Entanglement HCI The Next Wave? *ACM Transactions on Computer-Human Interaction* 27, 1 (2019), 1–27. https://doi.org/10.1145/3364998

Helena Anna Frijns, Oliver Schürer, and Sabine Theresia Koeszegi. 2021. Communication Models in Human–Robot Interaction: An Asymmetric MODel of ALterity in Human–Robot Interaction (AMODAL-HRI). *International Journal of Social Robotics* (2021), 28. https://doi.org/10.1007/ s12369-021-00785-7

Nora Fronemann, Kathrin Pollmann, and Wulf Loh. 2021. Should my robot know what's best for me? Human–robot interaction between user experience and ethical design. *AI & SOCIETY* (2021), 17. https://doi.org/10.1007/s00146-021-01210-3

Elizabeth Goodman, Erik Stolterman, and Ron Wakkary. 2011. Understanding interaction design practices. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Vancouver BC Canada). ACM, 1061–1070. https://doi.org/10.1145/1978942.1979100

Michael A. Goodrich and Alan C. Schultz. 2007. Human-Robot Interaction: A Survey. *Foundations and Trends® in Human-Computer Interaction* 1, 3 (2007), 203–275. https://doi.org/10.1561/1100000005

Kim Halskov and Caroline Lundqvist. 2021. Filtering and Informing the Design Space: Towards Design-Space Thinking. *ACM Transactions on Computer-Human Interaction* 28, 1 (2021), 1–28. https://doi.org/10.1145/3434462

Steve Harrison, Deborah Tatar, and Phoebe Sengers. 2007. The Three Paradigms of HCI. In *Alt. Chi. Session at the SIGCHI Conference on Human Factors in Computing Systems* (San Jose California USA). 1–18.

Frank Hegel. 2013. A Modular Interface Design to Indicate a Robot's Social Capabilities. In *ACHI 2013: The Sixth International Conference on Advances in Computer-Human Interactions*. 426–432.

Lars Erik Holmquist and Jodi Forlizzi. 2014. Introduction to Journal of Human-Robot Interaction Special Issue on Design. *Journal of Human-Robot Interaction* 3, 1 (2014), 3. https://doi.org/10.5898/ JHRI.3.1.Holmquist

Kasper Hornbæk and Antti Oulasvirta. 2017. What Is Interaction?. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver Colorado USA). ACM, 5040–5052. https://doi.org/10.1145/3025453.3025765

Kristina Höök and Jonas Löwgren. 2012. Strong concepts: Intermediate-level knowledge in interaction design research. *ACM Transactions on Computer-Human Interaction* 19, 3 (2012), 1–18. https://doi.org/10.1145/2362364.2362371

International Organization for Standardization (ISO). 2010. *ISO 9241-210:2010(en), Ergonomics of human-system interaction — Part 210: Human-centred design for interactive systems*. https://www.iso.org/obp/ui/#iso:std:iso:9241:-210:ed-1:v1:en

International Organization for Standardization (ISO). 2012. *ISO 8373:2012(en), Robots and robotic devices — Vocabulary*. https://www.iso.org/obp/ui/#iso:std:iso:8373:ed-2:v1:en

Jerry A Jacobs and Scott Frickel. 2009. Interdisciplinarity: A Critical Assessment. *Annual Review of Sociology* 35, 1 (2009), 43–65. https://doi.org/10.1146/annurev-soc-070308-115954

Paul Johannesson and Erik Perjons. 2014. *An Introduction to Design Science*. Springer International Publishing. https://doi.org/10.1007/978-3-319-10632-8

Ulla Johansson-Sköldberg, Jill Woodilla, and Mehves Çetinkaya. 2013. Design Thinking: Past, Present and Possible Futures. *Creativity and Innovation Management* 22, 2 (2013), 121–146. https://doi.org/10.1111/caim.12023

John Law. 1992. Notes on the theory of the actor-network: Ordering, strategy, and heterogeneity. *Systems Practice* 5, 4 (1992), 379–393. https://doi.org/10.1007/BF01059830

Paul M Leonardi. 2012. Materiality, Sociomateriality, and Socio-Technical Systems: What Do These Terms Mean? How Are They Different? Do We Need Them? In *Materiality and Organizing*, Paul M. Leonardi, Bonnie A. Nardi, and Jannis Kallinikos (Eds.). Oxford University Press, 24–48. https://doi.org/10.1093/acprof:oso/9780199664054.003.0002

Maria Luce Lupetti, Cristina Zaga, and Nazli Cila. 2020. Designerly HRI knowledge: Bridging HRI and Design Research. In *Proceedings of the 29th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN 2020).*

Maria Luce Lupetti, Cristina Zaga, and Nazli Cila. 2021. Designerly Ways of Knowing in HRI: Broadening the Scope of Design-oriented HRI Through the Concept of Intermediate-level Knowledge. In *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction* (Boulder CO USA). ACM, 389–398. https://doi.org/10.1145/3434073.3444668

Michal Luria, Marius Hoggenmüller, Wen-Ying Lee, Luke Hespanhol, Malte Jung, and Jodi Forlizzi. 2021. Research through Design Approaches in Human-Robot Interaction. In *Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction* (Boulder CO USA). ACM, 685–687. https://doi.org/10.1145/3434074.3444868

Jonas Löwgren. 2007. Pliability as an Experiential Quality: Exploring the Aesthetics of Interaction Design. *Artifact* 1, 2 (2007), 85–95. https://doi.org/10.1080/17493460600976165

Jeremy A Marvel, Shelly Bagchi, Megan Zimmerman, and Brian Antonishek. 2020. Towards Effective Interface Designs for Collaborative HRI in Manufacturing: Metrics and Measures. *ACM Transactions on Human-Robot Interaction* 9, 4 (2020), 1–55. https://doi.org/10.1145/3385009

Anja Naumann, Jörn Hurtienne, Johann Habakuk Israel, Carsten Mohs, Martin Christof Kindsmüller, Herbert A. Meyer, and Steffi Hußlein. 2007. Intuitive Use of User Interfaces: Defining a Vague Concept. In *Engineering Psychology and Cognitive Ergonomics*, Don Harris (Ed.). Vol. 4562. Springer Berlin Heidelberg, 128–136. https://doi.org/10.1007/978-3-540-73331-7_14 Series Title: Lecture Notes in Computer Science.

Antti Oulasvirta, Jussi P P Jokinen, and Andrew Howes. 2022. Computational Rationality as a Theory of Interaction. In *CHI Conference on Human Factors in Computing Systems* (New Orleans LA USA). ACM, 1–14. https://doi.org/10.1145/3491102.3517739

Andrea Owe and Seth D Baum. 2021. Moral consideration of nonhumans in the ethics of artificial intelligence. *AI and Ethics* (2021), 12. https://doi.org/10.1007/s43681-021-00065-0

Elisa Prati, Margherita Peruzzini, Marcello Pellicciari, and Roberto Raffaeli. 2021. How to include User eXperience in the design of Human-Robot Interaction. *Robotics and Computer-Integrated Manufacturing* 68 (2021), 13. https://doi.org/10.1016/j.rcim.2020.102072

Dimitrios Raptis, Jesper Kjeldskov, Mikael B Skov, and Jeni Paay. 2014. What is a Digital Ecology? Theoretical Foundations and a Unified Definition. *Australian Journal of Intelligent Information Processing Systems* (2014), 6.

Matt Ratto. 2016. Making at the end of nature. *Interactions* 23, 5 (2016), 26–35. https://doi.org/10.1145/2985851

Horst W J Rittel and Melvin M. Webber. 1973. Dilemmas in a General Theory of Planning. *Policy Sciences* 4, 2 (1973), 155–169. http://www.jstor.org/stable/4531523

Eike Schneiders, EunJeong Cheon, Jesper Kjeldskov, Matthias Rehm, and Mikael B. Skov. 2022. Non-Dyadic Interaction: A Literature Review of 15 Years of Human-Robot Interaction Conference Publications. *ACM Transactions on Human-Robot Interaction* 11, 2 (2022), 1–32. https://doi.org/10.1145/3488242

Herbert Alexander Simon. 2008. *The sciences of the artificial* (3rd ed.). MIT Press.

Gillian Crampton Smith. 2006. What Is Interaction Design? In *Designing Interactions*. The MIT Press,vii–xix.

Erik Stolterman. 2008. The Nature of Design Practice and Implications for Interaction Design Research. *International Journal of Design* 2, 1 (2008), 55–65.

JaYoung Sung, Rebecca E. Grinter, and Henrik I. Christensen. 2010. Domestic Robot Ecology: An Initial Framework to Unpack Long-Term Acceptance of Robots at Home. *International Journal of Social Robotics* 2, 4 (Dec. 2010), 417–429. https://doi.org/10.1007/s12369-010-0065-8

Rick Szostak, Claudio Gnoli, and María López-Huertas. 2016. *Interdisciplinary Knowledge Organization*. Springer International Publishing. https://doi.org/10.1007/978-3-319-30148-8

Núria Vallès-Peris and Miquel Domènech. 2021. Caring in the inbetween: a proposal to introduce responsible AI and robotics to healthcare. *AI & SOCIETY* (2021), 1–11. https://doi.org/10.1007/s00146-021-01330-w

Aimee Van Wynsberghe. 2016. Service robots, care ethics, and design. *Ethics and Information Technology* 18, 4 (2016), 311–321. https://doi.org/10.1007/s10676-016-9409-x

Aimee Van Wynsberghe and Shuhong Li. 2019. A paradigm shift for robot ethics: from HRI to human–robot–system interaction (HRSI). *Medicolegal and Bioethics* Volume 9 (2019), 11–21. https://doi.org/10.2147/MB.S160348

Peter-Paul Verbeek. 2008. Morality in Design, Design Ethics and the Morality of Technological Artifacts. In *Philosophy and Design*, P. E. Vermaas (Ed.). Springer, 91–103. https://classes.matthewjbrown.net/teaching-files/philtech/verbeek-design.pdf

Astrid Weiss, Regina Bernhaupt, Michael Lankes, and Manfred Tscheligi. 2009. The USUS Evaluation Framework for Human-Robot Interaction. In *AISB2009: proceedings of the symposium on new frontiers in human-robot interaction*, Vol. 4. 8.

Yueh-Hsuan Weng, Eunjoung Cheon, Phoebe Li, and Osamu Sakura. 2021. *1st Workshop on Design-Centered HRI and Governance*. https://krinuts7.wixsite.com/hri-design

Robert F Woodbury and Andrew L Burrow. 2006. Whither design space? *Artificial Intelligence for Engineering Design, Analysis and Manufacturing* 20, 2 (2006), 63–82. https://doi.org/10.1017/S0890060406060057

Albena Yaneva. 2009. Making the Social Hold: Towards an Actor-Network Theory of Design. *Design and Culture* 1, 3 (2009), 273–288. https://doi.org/10.1080/17547075.2009.11643291

J D Zamfirescu-Pereira, David Sirkin, David Goedicke, Ray Lc, Natalie Friedman, Ilan Mandel, Nikolas Martelaro, and Wendy Ju. 2021. Fake It to Make It: Exploratory Prototyping in HRI. In *Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction* (Boulder CO USA). ACM, 19–28. https://doi.org/10.1145/3434074.3446909

John Zimmerman, Erik Stolterman, and Jodi Forlizzi. 2010. An Analysis and Critique of Research through Design: towards a formalization of a research approach. In *DIS 2010* (Aarhus, Denmark). 310–319. https://doi.org/10.1145/1858171.1858228

Selma Šabanović. 2010. Robots in Society, Society in Robots: Mutual Shaping of Society and Technology as a Framework for Social Robot Design. *International Journal of Social Robotics* 2, 4 (2010), 439–450. https://doi.org/10.1007/s12369-010-0066-7

Selma Šabanović, Marek P. Michalowski, and Linda R. Caporeal. 2007. Making Friends: Building Social Robots through Interdisciplinary Collaboration. *AAAI Spring Symposium: Multidisciplinary Collaboration for Socially Assistive Robotics* (2007), 7.

Trust and Plausibility

Exploring the Situated Vulnerabilities of Robots for Interpersonal Trust in Human-Robot Interaction

Glenda Hannibal, Astrid Weiss

Abstract

The practical value of studying trust in human-robot interaction (HRI) rests on the assumption that people will, in the long-term, accept, interact, and collaborate more with robots that they trust or consider trustworthy. We propose in this book chapter to take our *event approach* to interpersonal trust in HRI and we argue why focusing on robot vulnerabilities will benefit current discussions on trust in robots and their perceived trustworthiness. On a theoretical level, we first argue that it is important to challenge the often negative view of the conceptual relationship between interpersonal trust and vulnerability in HRI as it has mainly comes to represent overexposure. Moreover, identifying robot-specific vulnerabilities is essential when exploring interpersonal trust in interactions between humans and robots (or HRI) because it overlaps but is not identical to those important to a human-centered perspective. To empirically explore robot vulnerabilities, we present the results of eight semi-structured expert interviews with experienced leaders in robotics. Based on these interviews, we identify the various robot vulnerabilities mentioned by the experts to present a systematic overview. Furthermore, we discuss how the experts interpreted the notion of vulnerability in relation to robots specifically and dive more into how malicious human behavior can be problematic when aiming to ensure mutual interpersonal trust in HRI. Moreover, we aim in this book chapter to lay down our motivation and arguments for why taking into account robot vulnerabilities provide a crucial and broader perspective on mutual trust in HRI, which is fundamental to strengthening interaction, collaboration and engagement between humans and robots.

Keywords

human-robot interaction, interpersonal trust, event approach, vulnerability, expert interviews, ethics, engagement

1 Introduction

The most common trust relation people have with artifacts and technologies is best described in terms of reliance and understood as a certain form of dependency. This dependency assumes that reliance on an inanimate object is necessary for the successful realization of some kind of plan given specific goals. Viewed as plan execution, trust as reliance mainly gets its value because of its ability to guide thoughts and actions from the perspective that seems reasonable given the means adopted to meet the concrete ends [Smith 2010; Alonso 2014]. Consequently, trust as reliance cannot be understood solely as something internal to the person trusting since it also depends on external conditions, which are laws of nature and the constraints of the specific design. Therefore, the main focus of trust as reliance is placed on making the interactions as smooth, efficient, and comfortable as possible in which artifacts or technologies are only to be considered instruments or tools to help people achieve their goals.

This instrumental view is the most traditional and widespread understanding of inanimate objects and also guides current understandings of robots [Coeckelbergh 2010a]. In robotics, *trust as reliance* is taken to mean that a person holds a predictive belief or assumption related to the performance of the robot given

https://doi.org/10.34727/2022/isbn.978-3-85448-052-5_2

the intended purpose and the predefined task. The performance of the robot then determines its trustworthiness and is considered important as it helps in establishing whether or not people are justified in trusting the robot. From this perspective, the robot's performance ensures that people can strike the right level of trust during interactions, collaboration, or engagement. Thus, an appropriate level of trust is treated as an indirect measure, which is later used to suggest specific design guidelines to prevent either under- or over-reliance [De Visser et al. 2020; Kok and Soh 2020].

The instrumental view has been significantly challenged with the recent aim in social robotics to make robots more socially capable and human-like (in regards to both physical appearance and style of behavior). Drawing on computational models of human cognition and social competence, "socially intelligent robots" [Breazeal 2001; Dautenhahn 1995] have built-in capacities to recognize and display cues for social interaction and communication. As such, they can behave and respond to people in a way they might interpret as intentional, influencing how people approach and treat robots. Similarly, endowing robots with anthropomorphic features only amplifies the tendency to perceive them as more human-like and is used as a deliberate design strategy to facilitate human-robot interaction (HRI) [Złotowski et al. 2016]. However, taking seriously human perception of robots as more socially capable and human-like also means that the current conceptualization of trust as reliance for HRI is no longer sufficient because it does not capture the additional social dimension of such interactions, which also extend to more ethical issues [Malle and Ullman 2021; Nyholm 2020]. Recent work on trust in HRI has attempted to adopt the notion of *interpersonal trust* to better study trust between humans and robots and uses this conceptualization as an explicit framework for the development of trustworthy robots [Lee et al. 2013; Ogawa et al. 2019].

In HRI research, speaking about interpersonal trust is taken to be unproblematic, and its meaning is connected to the observation that people seem to trust in robots and consider them trustworthy because of the assumed motives or intentions underlying their performance or actions due to their apparent agency. In such instances, speaking about interpersonal trust in HRI describes how people perceive robots as being concerned with their welfare, taking their views and personal interests into account, and working toward fair and unbiased outcomes. With these added social concerns, recent studies on trust in HRI have investigated how people attribute responsibility and blame to robots given an unfavorable outcome [Kaniarasu and Steinfeld 2014; Lei and Rau 2020]. These discussions bring forward the very ethical dimensions of human trust in robots and their perceived trustworthiness. The proposal to take into account the social and ethical dimensions of trust in HRI, through the application of the interpersonal trust no-

tion, is valuable as a first step to deepening our understanding of what happens in the interactions between humans and social robots. This work aids in recognizing that there is an added layer of complexity because it is no longer only a matter of performance but also about what follows from leveraging social rules and schemes to enhance the interaction. However, given the philosophical account of interpersonal trust compared to the technological advancement level in social robotics, Atkinson et al. asked the important question about the "appropriateness of using interpersonal trust as an analog for human-robot trust" [2012, p. 306]. They explained that making such an analogy has been argued as reasonable on the ground that some aspects of interpersonal trust also seem to be present in studies on HRI. However, not all fellow researchers are willing to draw such an analogy because of the lack of reciprocity in the interaction.

1.1. From Properties to the Event of Trust

What is interesting about this objection is that such concerns about reciprocity are a symptom of a more fundamental issue about the ontological status of the two kinds of agents. From a philosophical analysis, the issue of reciprocity touches upon the more basic ontological question of whether robots (as belonging to the class of inanimate objects) are of the *right kind* to be in the *category of objects that are appropriate targets of interpersonal trust* because their status as ontological equal to humans cannot be justified. Focusing on the ontological status of robots with a view to their properties is an intuitive and common way to reject robots as suitable objects of interpersonal trust. The needed argumentative step is to compare the relevant properties of robots with the criteria governing the category of objects that are appropriate targets of interpersonal trust established by "the 'official' philosophical inventory of things that are" [Loux and Crisp 2017, p. 13], which is also known as an ontology. The argumentative steps taken are of the general form:

- **Premise 1:** Having a certain property (P) is a necessary and sufficient criterion for belonging to the category of objects (C).
- **Premise 2:** All entities belonging to the category of objects (C) are appropriate targets of interpersonal trust.
- **Premise 3:** All entities that are part of the class inanimate objects (O) do not have the property (P).
- **Premise 4:** A robot (R) is a member of the class inanimate objects (O) .
- **Therefore:** A robot (R) does not belong to the category of objects (C) that are appropriate targets of interpersonal trust.

Although different suggestions can be made for the exact necessary and sufficient properties for members of the class of animate objects that belong to the category of objects, the notion of interpersonal trust cannot be directly applied without violating the basic requirements of both parties to be ontological equivalent as they share the same properties. However, using only the conceptualization of trust as mere reliance for the analysis of trust in HRI is undesirable because this conceptualization tends to significantly downplay the social and ethical dimensions that have already empirically proven to be relevant for human trust in robots and their perceived trustworthiness. Left unaddressed, speaking about interpersonal trust in the context of HRI forces complex metaphysical discussions about whether the relevant facts of ordinary language use in light of the truth of the relevant prephilosophical claims requires us to reevaluate whether the application of the interpersonal trust concept must be granted to robots or not. Therefore, speaking about interpersonal trust for HRI poses a challenge to the metaphysical theory of trust proposed by philosophers. A discussion that is not going to be settled easily or anytime soon. For those eager to empirically explore trust in HRI, a more pragmatic solution is required for this conceptual challenge. HRI researchers need to know the implications of such intricate philosophical discussion upon their work on trust in HRI that is motivated and held to the standard of empirical investigations. From this perspective, studies on trust in HRI must account for what happens despite better knowledge, especially in those instances where the apparent agency of robots is reflected in their use of language and their actions and behaviors toward robots that they trust or consider trustworthy.

We propose to shift the focus on trust in HRI away from only speaking about the properties of the parties involved in the interaction, but instead consider the *event* of interpersonal trust itself. This new outlook simply extends the unit of analysis beyond the identification of properties ascribed to either humans or robots to the circumstances where interpersonal trust happens. Such an approach considers not only *who* or *what* can be included in the category of objects that are appropriate targets of interpersonal trust, but also takes into account the *conditions* under which interpersonal trust occurs. Taking the study of trust in HRI to be an event, poses a new central question that is open also to empirical investigation: *Are the kind of interactions that occur between humans and robots some that could be labeled as interpersonal trust?* So even though humans and robots are still ontological of different kinds, this broader perspective permits the study of trust in HRI to consider the properties of the parties involved in the trust event without making these properties the dividing line of how we speak or consider the analysis of trust in interactions between humans and robots. From a methodological perspective, the important difference between the property and event approaches is that they operate with different criteria for the inclusion or exclusion of robots from the category of objects that are appropriate targets of interpersonal

trust. The property approach focuses on class membership of the right kind as the criterion. In contrast, the event approach considers the criterion of identity, which is to be understood as a principle stating the necessary and sufficient conditions for an event *E* and an event *E** to be identical [Bennett 1988]. We argue that our event approach for studying trust in HRI would serve the practical aim of bypassing the issues of ontological asymmetry between humans and robots while still being able to speak appropriately about interpersonal trust as the focus is now placed on the occurrence. We argue that the occurrence of interpersonal trust is bounded by the preconditions of trust.

Figure 1 Abramović and Ulay performing *Rest Energy* (1980). Courtesy of Marina Abramović and Sean Kelly Gallery, New York [Abramović 2016]. DACS 2016.

To get a quick idea about these preconditions, consider the famous and stunning art performance *Rest Energy* (1980) by Marina Abramović and Ulay that was first shown at ROSC'80 (see Figure 1). In this art performance, the two artists draw a bow and arrow to hold each other in suspension while small microphones placed under their shirts capture their accelerating heartbeats during the performance. A strong atmosphere of tension is created for around four minutes, as any wrong movement or a lapse of attention could be fatal for Abramović because the

arrow is pointing directly at her heart. While no longer in control of the situation, she is left exposed and Abramović later explained that the piece was “the ultimate portrait of trust.”[Abramović 2016, p. 255].

What this art performance can teach us is that trust is required under very specific circumstances:

1. When there is a possibility of harm (i.e., risk).
2. When there is a future-oriented likelihood of harm (i.e., uncertainty).
3. When this exposure leaves people vulnerable (i.e., vulnerability).

This art performance also illustrates that the relationship between trust and vulnerability is fundamental for understanding trusting relationships and that the occurrence of trust is a careful balance between the two parties involved as they try to prevent harm from happening. As we can see, Ulay tries not to harm (or even murder) Abramović while she does not want to be harmed even though the risk and uncertainty are evident to both of them.

1.2. Avoiding Overexposure

As Cipolla [2018] points out, there is often some reluctance to highlight this precondition when studying trust in relation to technology because “vulnerability is not usually interpreted positively, particularly when related to design or engineering” [Cipolla 2018, p. 113]. Mainly associated with overexposure to danger (i.e. risk) and unfamiliarity (i.e. uncertainty), discussions about vulnerability in regards to technology usage tends to be something that needs to be avoided, solved or explained away. Dagan et al. [2019] elaborate on this tendency in their motivation for the designing of the social wearable technology “True Colors”. They state that an explicit focus on vulnerability as a design value is rarely considered in the human-computer interaction (HCI) community, because technology is mainly seen as a tool empowering people to live a better, more pleasant, and safer life. If there are any vulnerabilities in sight, Dagan et al. [2019] continues, the developers often call for technological fixes or new innovations to solve these issues or reestablish a sense of security or protection. By characterizing this instrumental view on technology as a project of modernity, Coeckelbergh [2017] explain how the underlying assumption for the development and use of information and communication technology (ICT) reflects the agenda of vulnerability reduction. Coeckelbergh writes:

> “By means of using electronic devices, the Internet, and all kinds of ICT infrastructures we hope to become less vulnerable, to control risk. We hope to be less dependent on ’nature’, on ’the earth’, on our vulnerable bodies. We might

> even hope to liberate ourselves from a kind of Platonic dark cave where vulnerability and mortality reigns, and instead walk into the bright light of a new, invulnerable future" [2017, p. 344].

Therefore, it can be deduced from his account that the perception of technology as a form of remedy to all the possible harm of the world is a coping mechanism that does not recognize or leave any space for vulnerability. As such, it might not be too surprising that vulnerability, as an important theme for technology development and design, is rarely considered as something positive or worthwhile, unless it is merely to optimize our technological instruments and systems.

In HRI, focusing on vulnerability may also be considered problematic, but for a different reason. Through many years of ethnographic research into the way children and older adults respond and relate to robots developed to offer them companionship, Turkle [2011] warns us against how such new forms of technology can leave people very vulnerable. With her critical view on the promise of eliminating vulnerability through the reduction and simplicity of relationships by using robots to meet people's basic needs, the bad association of vulnerability with technology is now related to the danger of deception and its consequences on how people form emotional attachments. Turkle writes:

> "Technology is seductive when what it offers meets our human vulnerabilities. And as it turns out, we are very vulnerable indeed. We are lonely but fearful of intimacy. Digital connections and the sociable robot may offer the illusion of companionship without the demands of friendship" [2011, p. 1].

The strong message provided in this quote is that serious psychological harm can result from a false sense of intimacy when engaging with robots who seek to establish an emotional connection and that there is a level of enhancement involved in such kinds of interaction. The work by Turkle [2011] revolves to a large extent on presenting that the fascination with robots capable of imitating signs of care and love will eventually lead to unhealthy and unauthentic emotional attachments. This is because the possibility that such technologies offer is to spare people from the hardship and disappointment integral to developing deeper relationships with other people. By focusing on the vulnerability of people during HRI as a form of exploitation of both children and older adults who are in need of special care and love, several attempts have been made to better understand and discuss what can be done to avoid that people are potentially being deceived by robots [Sharkey and Sharkey 2020; Grodzinsky et al. 2015; Danaher 2020].

This rather gloomy outlook on the role vulnerability plays in our relation to robots is unfortunate when discussing trust in HRI. Because vulnerability is one of the preconditions of interpersonal trust, aiming to avoid vulnerability or trying to explain it away will paradoxically also undermine the demand for trust "in the ab-

sence of vulnerability trust is not required" [Misztal 2011, p. 117]. As she explain, if vulnerability is not of any concern in the first place there would be no need for anyone to trust in others because they would be able to meet their goals, needs, or gain prosperity free from the support or help of people. To live an invulnerable life would mean to be completely and utterly self-sufficient, a state that some might strive for and work hard to achieve; however, it is also still to be seen. This point was also well explained by Möllering when he wrote:

> "[...] in order to describe the typical experience of trust we often refer to the fact that actors trust *despite* their vulnerability and uncertainty, *although* they cannot be absolutely sure what will happen. They act as *if* the situation they face was unproblematic and, although they recognize their own limitations, they trust *nevertheless*" [2006, p. 6].

Central to our understanding of trust, as he shows, is that we are aware of our vulnerability but can interact and engage with the world anyway. We will argue that this is similar when we aim to understand and study interpersonal trust in HRI. Therefore, it is essential to challenge the rather negative view of the relationship between trust and vulnerability. Considering more recent studies on trust in HRI, it seems that there is already some empirical support for the consideration of vulnerability as something that is not only problematic, but could also support the interaction and engagement with robots.

1.3. Vulnerability and Trust in HRI[1]

The notion of vulnerability is similar to that of trust; it is very abstract, and its exact meaning can be hard to grasp. One way to understand what people have come to understand with vulnerability in the HRI community is to show that it have been operationalized. Several studies on trust in HRI currently take vulnerability to be some form of self-disclosure by a robot through verbal expressions and communication. Using such an understanding of vulnerability is very helpful when designing empirical studies because it is made less abstract (i.e. specific linguistic statements), which eventually render it more easily manipulated and measured. Consequently, all existing studies so far are designed to explore how expressions or utterances of vulnerability by a robot can influence human behavior or communication during HRI.

For example, Siino et al. [2008] found that a robot using a style of affective disclosure during a collaborative task in a repair scenario would result in people feeling less in control of their data but increased its like-ability. Even though, this study is not directly about trust in HRI, it is still interesting as the findings could

1 Subsection 1.3 has already been published as Hannibal [2021].

be understood as an expression of either human experience of vulnerability or perception of the robot being more vulnerable when reporting its affective state. In another example, Kaniarasu and Steinfeld [2014] were able to show that an utterance of self-blame by a robot during a collaborative task in a navigation scenario leads people to find it less trustworthy. As discussed by the authors, the tendency by people to view others negatively, who constantly make an apology for themselves despite their intention of being honest, is an effect seen in HRI that shed light on issues of distrust. However, some studies have suggested that robot self-disclosure can improve trust in HRI. Martelaro et al. [2016] found in their more recent study that, a simple robot expressing statements of vulnerability during a learning task in a tutorial scenario would result in a higher level of trust and sense of companionship. More interested in group dynamics, Sebo et al. [2018] found that when a robot during a collaborative task in a game scenario made vulnerable statements, the members would display a much higher level of engagement with it. Traeger et al. [2020] extended their work and found that the communication between the team members would improve, and their experience as part of the group would be seen positively when the robot provided statements of vulnerability. Reducing vulnerability in HRI to a form of self-disclosure is problematic in two ways.

First, operationalizing vulnerability only as the robot's behavior fails to recognize that vulnerability as a precondition of trust must always be interpreted and linked to the situatedness and temporality of the interaction. Thus, vulnerability is something that arises from the given circumstance in relation to a real and perceived vulnerability, depending on how the interaction plays out. Second, designing the vulnerability behavior of robots in the form of linguistic statements is a very narrow understanding of how robots could be vulnerable because it is a form of mimicking human vulnerability. Considering the literature so far on robot failures (e.g., [Salem et al. 2015; Ragni et al. 2016; Honig and Oron-Gilad 2018]) and cybersecurity in robotics (e.g., [Clark et al. 2017; Miller et al. 2018]), the way in which robots can be vulnerable only partially overlaps with human vulnerabilities. In other words, given that robots are ontologically of a different kind, they have their own specific types of vulnerabilities. Hence, systematically identifying these robot-specific vulnerabilities is in fact equally important to the identification of human vulnerabilities when exploring trust in HRI. As such, a gap in the current research landscape has been identified, which serves as the motivation for the expert interviews presented in the next section. Moreover, reducing vulnerability only to a property of the robot's behavior fails to recognize that vulnerability, as a precondition of trust, must always be interpreted and linked to the specific situation or moment in time. As we wish to highlight also in the later discussion about the expert interview results, it is important to include the insight that vulnerability is relational in the research on trust in HRI, because it is highly sensitive to the

ongoing and ever-changing relationship between humans and robots during interaction.

2 Expert Interviews[2]

Given these theoretical perspective, we set out to explore the aspect of vulnerability as a precondition of trust in HRI by gathering knowledge about the possible robot vulnerabilities. Guiding this research with the question of *in which way robots could be considered vulnerable?*, we decided to conduct semi-structured expert interviews with experienced and leading roboticists.

2.1. Methodology

The method for conducting expert interviews is suitable for getting a more systematic overview of knowledge within certain domains, which experts have spent many years achieving through their professional training or experience [Meuser and Nagel 2009]. For this research, expert interviews are helpful in the initial stage of identifying the possible vulnerabilities of robots. Not only do robotics experts have an extensive knowledge about the technical challenges of developing robots, they can also provide insights into what types of vulnerabilities are common across various domains of application.

On a methodological level, using expert interviews is important because of the ontological status of robots. First, given that robots do not have an inner life that connects feelings of vulnerability to higher mental states or experiences, their particular vulnerabilities can only be studied from a third-person perspective. To paraphrase Bruno Latour, whose words about scientific facts are equally relevant to this discussions, expert interviews are required because robots cannot “speak for themselves” [Latour 1993, p. 29]. Thus, we take the specialized knowledge of roboticists as a vehicle for giving expression to the specific vulnerabilities of robots.

2.2. Procedure

Over the period of nine months, we conducted in total eight semi-structured expert interviews. The purposeful sampling method [Patton 2015] was used to recruit the experts with the following selection criteria (see e.g., Table 1 for a quick overview of how the different expertise was divided among the different experts):

2 Section 2 of this book chapter has already been published as Hannibal [2021].

1. Having a disciplinary background in robotics.

2. Work experience in HRI or social robotics.

3. Research interest on the topic of trust.

To address the research question, it was enough if an expert would only fulfill one of the three criteria while ideally they would cover all of them.

Experts	**ID**	**Country**	**Expertise**
Justus Piater	Exp_JP	AT	computer vision, ML, robotics
Allan Wagner	Exp_AW	USA	AI, robotics, HRI, robot ethics, trust
Marc Hanheide	Exp_MH	UK	AI, robotics, HRI, social robotics
–	Exp_XX	–	social robotics, HRI, AI, trust
Birgit Graf	Exp_BG	DE	HRI, service robotics, applications
Kristin Schaefer-Lay	Exp_KS	USA	robotics, HRI, teams, trust
Michael Zillich	Exp_MZ	AT	computer vision, robotics, HRI
Paul Robinette	Exp_PR	USA	robotics, HRI, trust

Table 1 Overview of the experts and the used selection criteria for their inclusion.

All experts were contacted via email with an invitation to participate, which also contained more background information and the purpose of the interview. After indicating their willingness to participate in the interview, all experts were asked to sign a consent form that was sent to them in advance. The consent form clearly stated what their participation involved, their rights, and the data protection requirements set by the university. Each expert interview was conducted in English, audio recorded, and took about 30-40 minutes.

In the first part of the interview, all experts were given an opportunity to introduce themselves (i.e., “Could you tell me about your recent projects and main research interest?”). This information was needed to contextualize their disciplinary background and role as experts (see e.g., Section 2.3). Then five additional questions were asked to guide the semi-structured interviews:

- What do you consider as future application scenarios for agent-like robotic systems?
- Given your research background, how and when can an agent-like robotic systems be said to be vulnerable?

- Given your considerations of system-centered vulnerabilities, could you please rank or order them according to their importance?
- From your point of view, who would be disadvantaged if these vulnerabilities are left unaddressed?
- Considering cutting-edge technical knowledge used to develop agent-like robotic systems today, what has to be done to make agent-like robotic systems less vulnerable in your opinion?

After finishing the interview, all experts had the opportunity to give feedback and were again informed about their rights as participants.

2.3. Ethics

To ensure the protection and integrity of the experts participating, we generally followed the four-fold strategy suggested by Flick [2009]: (1) ensure voluntary consent by the participants in advance, based on sufficient and adequate information about the research project and its aim, (2) avoid causing any unnecessary harm to the participants in the process of collecting data, (3) do justice to the participants when analyzing and interpreting the collected data, and (4) guarantee the confidentiality and anonymity of all the participants when writing down and presenting the results and findings. However, given the nature of expert interviews, we excluded principle 4 for the informed consent of the experts, as it stated in the consent that we could use the name, professional title and affiliation for the purpose of direct quotations. Only one of the experts wished to remain anonymous, who we have given the expert code, Exp_XX.

On a practical level, it is important to mention that there was no official ethics board at TU Wien that was in charge of providing a standardized procedure for ethical approval of the expert interviews at the time when they were conducted. Only since 2020 has TU Wien been testing a concept of a Research Ethics Committee (Pilot REC) based on peer review to ensure a future procedure for basic standards of research ethics. However, we did our best to compensate for the lack of ethical approval because we were in contact with Dr. Marjo Rauhala about the development of the expert interviews. Since Dr. Rauhala supports all researchers at TU Wien daily with the identification of questions regarding research ethics in the role as the leader of the service unit of Responsible Research Practices[3], we received some feedback on the project description and consent form provided to the experts, so they would live up to basic standards for good research practice. For guidance about how to follow the EU regulations of GDPR, the third author

3 For more information about the service unit of Responsible Research Practices at TU Wien, we suggest visiting their website: https://www.tuwien.at/en/research/rti-support/ responsible-research-practices

ensured a check since he is in the role of Data Protection Coordinator at the Faculty of Informatics, TU Wien. This information was also provided on the consent forms that the experts were asked to sign to prepare for their interviews.

2.4. Analysis

After collecting all the expert interviews, the audio recordings were transcribed verbatim with the spoken word as the only focus [McLellan et al. 2003]. The first author coded the interviews solely using in vivo coding to summarize the exact wording, terminology, and formulations used by the expert. After a few coding cycles, related codes were then lumped into overall 13 different categories based on content and meaning similarity [Miles et al. 2020]. The decision on which category labels to use was also guided by prior classification of potential system-centered vulnerabilities as reported in previous literature on robot failures [Ragni et al. 2016; Honig and Oron-Gilad 2018] and cybersecurity in robotics [Clark et al. 2017; Miller et al. 2018]. We used a thematic analysis [Braun et al. 2019] to identify the common themes across the expert interviews. All coding, categorization, and thematic analysis of the expert interviews were done electronically using MAXQDA[4].

Theme	Category
(T1) Embodiment	(C1) Mechanical
	(C2) Sensory
	(C3) Functional
	(C4) Security
(T2) Processing	(C5) Understanding
	(C6) Learning
	(C7) Decision-making
(T3) People	(C8) Obstacle
	(C9) Perspective-taking
	(C10) Malicious
(T4) Setting	(C11) Infrastructure
	(C12) Environment
	(C13) Time

Table 2 List of the different categories and themes identified during the coding and analysis of the expert interviews.

4 Due to the COVID-19 outbreak in March 2020, all but the first expert interview were conducted online using the Skype platform.

From the analysis of the expert interviews, we were able to identify in total 13 categories of vulnerability that were then grouped into four different themes (see e.g. Table 2). Next, we provide a short description of each theme and offer few examples of how they were supported by different experts by drawing on their own wording, terminology, and formulations to summarize their main points.

2.4.1. Embodiment (T1)

Since robots are navigating and interacting with people in the real world, they have on the most basic level what experts Exp_JP and Exp_KS referred to as "physical vulnerability". Under this theme, we collected all the various vulnerabilities related to robots in the sense that they could be "fragile" (Exp_JP), "damaged" (Exp_KS), "worn down" (Exp_BG), "hacked" (Exp_XX), or "break down" (Exp_AW). As such, the aim of this theme was to highlight that regarding their various mechanical (C1), sensory (C2), functional (C3), and security (C4) aspects, robots can be exposed because their required embodiment creates tangible vulnerabilities.

2.4.2. Processing (T2)

On a more abstract level, but still related to the functioning of robots, the next theme is related to their ability to handle and use the information they get from the surroundings for understanding (C5), learning (C6), and decision-making (C7) as Exp_JP mentioned that "softwares are also vulnerable". Central to this theme are the different robot vulnerabilities that arise because they "lack a conceptual framework that allows them to understand what is going on in the world" (Exp_JP), could be "learning the wrong thing" (Exp_MH), or could "make decisions when they do not have all of the information" (Exp_XX). Thus, these kinds of robot vulnerabilities are mainly to be understood as a form of exposure in the sense of inadequate, misinformed, and hasty reasoning that eventually guides their behavior.

2.4.3. People (T3)

Moving on to those aspects that are more external to the robot, the next theme relates specifically to the action or behavior of the people interacting with them and that would have a direct effect on their level of exposure. For some of the experts, the robot vulnerabilities were not simply a matter of people sometimes hindering task completion by the robot because "the human does not move so the robot has to turn" (Exp_PR) or that the limited "understanding in humans how the

robots see the world" (Exp_MH) would result in robots getting into various accidents. In some cases people would in fact be downright "malicious" (Exp_PR) as they would intentionally engage in "abusive, aggressive behavior towards robots in the public" (Exp_MH). Thus, this theme intends to show that vulnerabilities are closely linked to both the unintentional and intentional conduct of the human counterpart because people expect that robots can easily deal with constantly moving obstacles (C8), fail to understand or take into account the perspectives of robots (C9) that leads to hazardous situations, and assume that mistreating robots by participating in malicious (C10) activities is unproblematic.

2.4.4. Setting (T4)

In the last theme, we collected the robot vulnerabilities mentioned by the experts, which relate to the framing or backdrop against which the interaction between humans and robots unfolds. In this theme, the often hidden technological and bureaucratic infrastructure (C11) was stressed because getting robots to properly function in real world scenarios often requires "ten engineers standing around" (Exp_BG) or "getting safety certificates" (Exp_MZ) to ensure robots could leave the laboratory and enter the market. Even when being tested for application, robots regularly get challenged when having to navigate in an environment (C12) designed for humans, which Exp_BG identified when she explained that "sometimes the corridor was simply too narrow" (Exp_BG) or that people would constantly be "moving stuff around". Time (C13) was also considered important given that according to Exp_KS there is a difference between those robot vulnerabilities that only show in the long-term compared to those "that happen right away". In the view of Exp_MH, the aspect of time might also be crucial in understanding why people "like to mess around" - because new and short encounters instigate a "novelty effect".

2.5. Discussion

There are several points to consider for discussing the results, which we will present in this section while relating them to existing literature in HRI and other relevant discussions.

2.5.1. Interpretation of Vulnerability

As expected, some of the experts would comment on how to interpret the notion of vulnerability in relation to robots. For example, Exp_PR considered how to understand robot vulnerability in light of how they are often portrayed in the media

and pop-culture. He noted that while people always see in movies that "robots are super strong and super fast and everything" this is far from the case because in "the real world they cannot get over a single step or they think that a bush is an obstacle that cannot be driven or something". Thus, Exp_PR concludes, that robots are "already pretty vulnerable in the real world" compared to the impression that the general public might have. This point is closely related to debates in HRI about managing public expectations regarding the robot's capabilities. Known by now as the "expectation gap" [de Graaf et al. 2016; Kwon et al. 2016], it is also highly relevant and recently linked to discussions regarding trust in HRI, as this gap could result in unwanted disappointment and even instigate fear [Malle et al. 2020].

More concerned with some conceptual challenges, Exp_AW expressed difficulties with speaking about robot vulnerabilities when saying that "vulnerability is just not a topic that's really very well suited for robots" because in his view using this notion would suggest that robots have some kind of volition or intentionality. Exp_AW further explained how this issue made him hesitate in using the common definition of trust by Mayer et al. [1995] and instead turned toward a "definition that involved risk", which is more practical and widespread in robotics since it is easier to operationalize. Another similar reflection was made by Exp_BG who said that "it's really hard to think about vulnerable in the sense of the robot because for me it's an attribute that's so human". Based on her more technical perspective, she then suggested reformulating the relevant aspect of considering robot vulnerability in terms of "situations where the robot could run into problems". This conceptual tension when studying trust in HRI has previously been identified by Malle and Ullman [2021] and it is still an open question whether human-robot trust necessarily comes with a feeling of vulnerability, which is a characteristic of human trust.

According to Exp_KS, such discussion must consider that speaking about robot vulnerabilities also contains a normative dimension because people in different contexts might need to ask themselves critically, "how vulnerable do we need to be to the system, how vulnerable does the system need to be to me?". She elaborates on this point by saying that robot vulnerabilities in a military context must always be avoided, whereas it might be useful in healthcare for building trust between people and robots. Questions about when and for what reasons robot vulnerabilities might be desirable or not are important to discussions about trust in HRI because the mere presence of a robot perceived as vulnerable can in fact influence human group dynamics for the better [Traeger et al. 2020].

2.5.2. Ethical Dimensions

From the expert interviews, it turned out that the theme of people (T3) ranked as the second most mentioned robot vulnerability despite different domains of application (coded 57 times). Especially the challenge of malicious humans was mentioned by several experts, who noted that people would intentionally be “kicking”,“pushing”, “hitting”, and “attacking” robots, which adds to previous HRI literature reporting how both adults and children would not shy away from such behavior [Scheeff et al. 2002; Brscić et al. 2015; Nomura et al. 2016]. This abusive behavior toward robots will only grow with their increasing application in public spaces, which according to Exp_KS is problematic for trust in HRI because “it will become an issue for their operation”. Given that the success or failure of a given task in fact depends on some level of mutual trust in HRI, it is relevant to study not only whether people can trust robots, but also whether robots can trust people [Vinanzi et al. 2019].

The necessity of mutual trust in HRI for task completion and collaboration requires a broader discussion about how to deal with human abusive behavior toward robots, and this challenge has already been recognized as an ethical dimension of HRI [Whitby 2008].

From a critical analysis of previous attempts in philosophy to account for trust that mainly originated from a liberal tradition, Baier [1986] argued that the significance of trust for thriving must be examined from a moral point of view. From her perspective, it is a bad starting point for any understanding of trust pertinent to human social life to consider it as some form of contract established between two equal parties in terms of power and capabilities. From her careful observation of interpersonal relationships of all kinds where cooperation and care are cardinal, she recognizes that some of them are fundamentally unequal and sometimes not even voluntary, which severely challenges the liberal ideal of the conditions of trust. Based on this insight, Baier [1986] propose instead to take trust as a form of reliance on others to act out of goodwill toward oneself. This so-called goodwill account of trust is essential in stressing the close connection between interpersonal trust with moral obligations and is one of the first views on trust that goes beyond reliance.

However, debates about mutual trust rooted in a liberal tradition have become challenging for HRI because they presume that the two parties stand to each other in an equal moral and power relation [Faulkner and Simpson 2017]. The acknowledgment of robot vulnerability in relation to their human counterpart is ethically problematic as they can at most be considered “moral patients” [Coeckelbergh 2018], and they do not have a choice whether or not to engage in the interaction [Baier 1986].

Considering both the limited moral standing of robots and the inequality of power in HRI, we agree with Tolmeijer et al. [2020] that future work needs to focus more on developing concrete trust-repair strategies for what they refer to as “user failure” to mitigate robot vulnerabilities resulting from abusive behavior. From their main focus on interaction design strategies for mutual trust in HRI, they have suggested that robots could use methods of apology, showing emotions, and involving authority figures. More concerned with ethical and legal strategies, debates in philosophical circles have been revolving around granting some form of “robot rights” [Coeckelbergh 2010b; Gunkel 2018], which is a rather controversial suggestion [Tavani 2018].

3 Relational Dimension of Vulnerability

Throughout his work on developing a normative anthropology of vulnerability, Coeckelbergh [2013] draws on the traditions of phenomenology and pragmatism for analyzing vulnerability in relation to technology, as an alternative to the more classical scientific approach. As he writes, the understanding that the classical sciences brings to the foreground of the discussion is one where “vulnerability appears as an objective, essential feature of human nature, and the vulnerability of people is studied in an objectivist way” [Coeckelbergh 2013, p. 38-39]. From this perspective, he continues, vulnerability is something external to people, which can be evaluated from a third-person point of view. Vulnerably is thereby characterized in objective terms; is vulnerability *real* compared to the possible risk and uncertainty posed by a threat to the livelihood or well-being of people. In this sense, the individual experience of being vulnerable is not considered or at least something that can be managed when understood properly. As Coeckelbergh [2013] explains, those who do in fact speak about vulnerability as tied to the subjective feelings or emotions of people still presuppose that the perception of being vulnerable is seen in the light of an objective standard. Taking the complete opposite view, is to consider vulnerability only as subjective where the first-person perspective is in focus, how the “I” (or individual) comes to experience the vulnerability. However, he argues that this view is also problematic because it does not acknowledge that the subjective experience of vulnerability is influenced by the surroundings and conditions people find themselves in. Vulnerability is connected to the way people interact and engage with the world, which contains both risk and uncertainty as part of daily life. Thus, Coeckelbergh [2013] aims to challenge this overall idea of the object-subject dichotomy to our understanding of vulnerability ingrained in the Western thought. As a way out of this dualistic view on vulnerability, he proposes to shift the focus on how vulnerability emerges out of this tension so that it “[...] is neither a feature of the world (an objective,

external state of affairs) nor something that we create or perceive (a subjective construction by the mind, an internal matter), but is constituted in the subject-object relation" [Coeckelbergh 2013, p. 43].

From this critical discussion, Coeckelbergh [2013] elaborates on what he means when he takes vulnerability to be relational, that closely connected with the notion of engagement. He states that vulnerability arises from or comes into view only within the relation that manifests when people engage with the world. It is nothing that already belongs to people or the world in advance, but something that unfolds in that meeting. Following this understanding of vulnerability as something emergent during the interaction is also relevant to the way it is possible to think about vulnerability for studies on trust in HRI. Given that vulnerability fundamentally emerges from the interaction or engagement between humans and robots, it would be a mistake to reduce it to being a property of the robots nor of the perceptions people have, as reported in from previous work. Rather, it is something that must be located in the event of a meeting. As relational vulnerability in HRI, we can take the co-constitution of vulnerability as a result of both the human and robot who are coming into interaction or engagement. While Coeckelbergh puts a lot of effort into stressing the value of this analysis because it makes room for the existential dimensions of a "vulnerable being" [2013, p. 44], we argue that the more important point he makes, and the most relevant for the HRI community, is that it also enables us to see vulnerability as a process; vulnerability is continuously ongoing. Since vulnerability is relational in terms of interaction and engagement, it also means that it is always in the making. Coeckelbergh makes this point clear when he writes:

> "Vulnerability is not merely passive. To understand vulnerability as something entirely passive would be to turn the human being into an object once again or a *property* of that object. But *openness* does not mean passivity, and vulnerability is not merely a characteristic of our body or our mind. We are not vulnerable in the way a building or a bridge is vulnerable. Rather, we *make* ourselves vulnerable; we put ourselves at risk, by our mental and physical actions. We eat, we travel, we work, we love, we hope, and these actions make us vulnerable. Vulnerability, therefore, is not a property of the human person but a feature of the relation between us and the world. It is a feature of our way of being (in the world) and a way of existing" [2013, p. 44].

Translating this insight into the context of trust in HRI, we can say that it is possible to consider vulnerability as a result of the exchange between the human and robot that always occur. Although robots are of a completely different kind than humans, we believe that this does not hinder the recognition that they play their own important role in the creation of vulnerability. Just as anything else in the world, which confronts people as part of their everyday life, our meeting with

robots can potential shape the way we come to experience and understand our vulnerability through encounters. This is similar to how robots can be considered vulnerable in the meeting with people. They are also affected by the actions and behaviors of humans, even though the issues that robots face from such meetings might not have the same existential consequences. However, there are potential risks and uncertainties that robots face when navigating in human spaces, which render them vulnerable and thus bring the theme of trust as bidirectional into the discussion.

4 Conclusion

In this book chapter, we have considered some theoretical and empirical work in deepening an understanding of interpersonal trust in HRI. First, we considered how trust had been understood in the context of HRI on a conceptual level, leading to deeper philosophical questions about the metaphysics of taking trust to be an event rather than a property as a way to highlight vulnerability as one of the preconditions less explored. Then, we then presented the results of eight expert interviews that aimed to explore how robots could be said to be vulnerable in interactions requiring trust. Based on the systematic overview, we discussed how robot vulnerability is challenging our conceptual associations and how such a stance leads to broader social and ethical discussions on trust in HRI, where mutual trust is essential in strengthening the interaction or collaboration. Finally, we reflected on how the current shift toward vulnerability as an emergent aspect of mutual trust in HRI aligns with a general view on how interpersonal trust is always a result of the ongoing exchange between humans and robots, even though they are of ontological different kinds.

In summary, our book chapter presents an interdisciplinary perspective on the analysis of trust for current HRI research. Although there are still many open questions to be addressed and further empirical work to be carried out, we believe that the initial steps have been taken toward new directions of understanding and studying trust in HRI. Furthermore, our work is also helpful in fostering a stronger dialog about how to combine both theoretical and empirical perspectives on the complex way of recognizing robot vulnerabilities that can support trust in HRI.

Bibliography

Marina Abramović. 2016. *Walk Through Walls: A Memoir.* Crown Archetype, New York (NY), USA. 304 pages.

Facundo M Alonso. 2014. What is reliance? *Canadian Journal of Philosophy* 44, 2 (4 2014), 163–183.

David Atkinson, Peter Hancock, Robert R Hoffman, John D Lee, Ericka Rovira, Charlene Stokes, and Alan R Wagner. 2012. Trust in Computers and Robots: The Uses and Boundaries of the Analogy to Interpersonal Trust. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 56, 1, 303–307.

Annette Baier. 1986. Trust and Antitrust. *Ethics* 96, 2 (1986), 231–260.

Jonathan Bennett. 1988. *Events and Their Names.* Vol. 28. Clarendon Press, Oxford, UK. 243 pages. https://doi.org/10.2307/2215925

Virginia Braun, Victoria Clarke, Nikki Hayfield, and Gareth Terry. 2019. Thematic Analysis. In *Handbook of Research Methods in Health Social Sciences*, P. Liamputtong (Ed.). Springer Nature Singapore Pte Ltd., Singapore, Chapter 48, 843–860.

Cynthia Breazeal. 2001. Socially Intelligent Robots: research, development, and applications. In *IEEE International Conference on Systems, Man and Cybernetics,* Vol. 4. IEEE, Tucson (AZ), USA, 2121–2126.

Drazen Brscić, Hiroyuki Kidokoro, Yoshitaka Suehiro, and Takayuki Kanda. 2015. Escaping from Children's Abuse of Social Robots. In *Proceedings of the 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI'15)*. ACM, Portland (OR), USA, 59–66.

Carla Cipolla. 2018. Designing for Vulnerability: Interpersonal Relations and Design. *She Ji: The Journal of Design, Economics, and Innovation* 4, 1 (2018), 111–122.

George W Clark, Michael V Doran, and Todd R Andel. 2017. Cybersecurity issues in robotics. In *Proceedings of the IEEE Conference on Cognitive and Computational Aspects of Situation Management (CogSIMA)*. IEEE, Savannah (GA), USA, 1–5.

Mark Coeckelbergh. 2010a. Artificial Companions: Empathy and Vulnerability Mirroring in Human-Robot Relations. *Studies in Ethics, Law, and Technology* 4, 3 (2010), Article 2.

Mark Coeckelbergh. 2010b. Robot rights? Towards a social-relational justification of moral consideration. *Ethics and Information Technology* 12, 3 (2010), 209–221.

Mark Coeckelbergh. 2013. *Human being @ risk: Enhancement, technology, and the evaluation of vulnerability transformations*. Springer (Science & Business Media), Berlin. 372 pages.

Mark Coeckelbergh. 2017. The Art of Living with ICTs: The Ethics-Aesthetics of Vulnerability Coping and Its Implications for Understanding and Evaluating ICT Cultures. *Foundations of Science* 22, 2 (2017), 339–348.

Mark Coeckelbergh. 2018. Why care about robots? Empathy, moral standing, and the language of suffering. *Kairos. Journal of Philosophy & Science* 20, 1 (2018), 141–158.

Ella Dagan, Elena Márquez Segura, Ferran Altarriba Bertran, Miguel Flores, and Katherine Isbister. 2019. Designing 'True Colors': A Social Wearable that Affords Vulnerability. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems.* ACM, Glasgow, Scotland, 1–14. https://doi.org/10.1145/3290605.3300263

John Danaher. 2020. Robot Betrayal: a guide to the ethics of robotic deception. *Ethics and Information Technology* 22 (2020), 117–128.

Kerstin Dautenhahn. 1995. Getting to Know Each Other - Artificial Social Intelligence for Autonomous Robots. *Robotics and Autonomous Systems* 16, 2-4 (1995), 333–356.

Maartje M A de Graaf, Somaya Ben Allouch, and Jan A G M van Dijk. 2016. Long-term evaluation of a social robot in real homes. *Interaction Studies* 17, 3 (12 2016), 461–490.

Ewart J De Visser, Marieke, M M Peeters, Malte, F Jung, Spencer Kohn, Tyler, H Shaw, Richard Pak, and Mark A Neerincx. 2020. Towards a Theory of Longitudinal Trust Calibration in Human-Robot Teams. *International Journal of Social Robotics* 12 (2020), 459–478.

Paul Faulkner and Thomas Simpson. 2017. Introduction. In *The Philosophy of Trust,* P Faulkner, T Simpson (Eds.). Oxford University Press, Oxford, UK. 3–14 pages.

Uwe Flick. 2009. *An Introduction to Qualitative Research* (4th ed.). SAGE Publications Ltd., London. 504 pages.

Frances S Grodzinsky, Keith W Miller, and Marty J Wolf. 2015. Developing Automated Deceptions and the Impact on Trust. *Philosophy & Technology* 28, 1 (3 2015), 91–105.

David J Gunkel. 2018. The Other Question: Can and Should Robots Have Rights? *Ethics and Information Technology* 20, 2 (2018), 87–99.

Glenda Hannibal. 2021. Focusing on the Vulnerabilities of Robots through Expert Interviews for Trust in Human-Robot Interaction. In *16th ACM/IEEE International Conference on Human-Robot Interaction.* ACM, Boulder, CO, 288–293.

Shanee Honig and Tal Oron-Gilad. 2018. Understanding and Resolving Failures in Human-Robot Interaction: Literature Review and Model Development. *Frontiers in Psychology* 9, 861 (2018). https://doi.org/10.3389/FPSYG.2018.00861

Poornima Kaniarasu and Aaron M. Steinfeld. 2014. Effects of blame on trust in human robot interaction. In *The 23rd IEEE International Symposium on Robot and Human Interactive Communication.* IEEE, Edinburgh, UK, 850–855.

Bing Cai Kok and Harold Soh. 2020. Trust in Robots: Challenges and Opportunities. *Current Robotics Reports* 1, 4 (12 2020), 297–309.

Minae Kwon, Malte F. Jung, and Ross A. Knepper. 2016. Human expectations of social robots. In *Proceedings of the 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI'16)*. IEEE, Christchurch, New Zealand, 463–464.

Bruno Latour. 1993. *We Have Never Been Modern.* Harvard University Press, Cambridge (MA), USA. 168 pages.

Jin Joo Lee, W. Bradley Knox, Jolie B. Wormwood, Cynthia Breazeal, and David DeSteno. 2013. Computationally modeling interpersonal trust. *Frontiers in Psychology* 4 (12 2013), 893. https://doi.org/10.3389/fpsyg.2013.00893

Xin Lei and Pei-Luen Patrick Rau. 2020. Should I Blame the Human or the Robot? Attribution Within a Human–Robot Group. *International Journal of Social Robotics* (4 2020), 1–15. https://doi.org/10.1007/s12369-020-00645-w

Michael J Loux and Thomas M Crisp. 2017. *Metaphysics: A contemporary introduction* (4 ed.).Routledge, New York (NY), USA. 1–356 pages.

Bertram F Malle, Kerstin Fischer, James E Young, Ajung Moon, and Emily C Collins. 2020. Trust and the discrepancy between expectations and actual capabilities of social robots. In *Human-robot interaction: Control, analysis, and design,* D. Zhang and B. Wei (Eds.). Cambridge Scholars Publishing, New York, NY, USA, Chapter 1, 1–23.

Bertram F Malle and Daniel Ullman. 2021. A Multi-Dimensional Conception and Measure of Human-Robot Trust. In *Trust in Human-Robot Interaction: Research and Applications,* C. S. Nam and J. B. Lyons (Eds.). Academic Press, London, UK, Chapter 1, 3–25.

Nikolas Martelaro, Victoria C Nneji, Wendy Ju, and Pamela Hinds. 2016. Tell me more: Designing HRI to encourage more trust, disclosure, and companionship. In *11th ACM/IEEE International Conference on Human-Robot Interaction (HRI).* IEEE, Christchurch, New Zealand, 181–188.

Roger C Mayer, James H Davis, and F David Schoorman. 1995. An Integrative Model of Organizational Trust. *The Academy of Management Review* 20, 3 (1995), 709–734.

Eleanor McLellan, Kathleen M. MacQueen, and Judith L. Neidig. 2003. Beyond the Qualitative Interview: Data Preparation and Transcription. *Field Methods* 15, 1 (2 2003), 63–84.

Michael Meuser and Ulrike Nagel. 2009. The Expert Interview and Changes in Knowledge Production. In *Interviewing Experts* (1st ed.), Alexander Bogner, Beate Littig, and Wolfgang Menz (Eds.). Palgrave Macmillan, Hampshire, UK, Chapter 1, 281.

Matthew B Miles, A Michael Huberman, and Johnny Saldaña. 2020. *Qualitative Data Analysis: A Methods Sourcebook* (4 ed.). SAGE Publications Ltd., Thousand Oaks, CA. 381 pages.

Justin Miller, Andrew B Williams, and Debbie Perouli. 2018. A Case Study on the Cybersecurity of Social Robots. In *Proceedings of 13th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. ACM, New York (NY), USA, 195–196.

Barbara A Misztal. 2011. *The Challenges of Vulnerability: In Search of Strategies for a Less Vulnerable Social Life.* Palgrave Macmillan, Hampshire, UK. 272 pages.

Guido Möllering. 2006. *Trust: Reason, Routine, Reflexivity.* Emerald Group Publiching Limited, Bingley, UK. 244 pages.

Tatsuya Nomura, Takayuki Kanda, Hiroyoshi Kidokoro, Yoshitaka Suehiro, and Sachie Yamada. 2016. Why do children abuse robots? *Interaction Studies* 17, 3 (12 2016), 347–369.

Sven Nyholm. 2020. *Humans and Robots: Ethics, Agency, and Anthropomorphism.* Rowman & Littlefield International, London, UK. 236 pages.

Rui Ogawa, Sung Park, and Hiroyuki Umemuro. 2019. How Humans Develop Trust in Communication Robots: A Phased Model Based on Interpersonal Trust. In *Proceedings of the 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI).* IEEE, Daegu, South Korea, 606–607.

Michael Quinn Patton. 2015. *Qualitative Research & Evaluation Methods: Integrating Theory and Practice* (4 ed.). SAGE Publications, Inc., Thousand Oaks, CA. 806 pages.

Marco Ragni, Andrey Rudenko, Barbara Kuhnert, and Kai O Arras. 2016. Errare humanum est: Erroneous robots in human-robot interaction. In *The 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, New York (NY), USA, 501–506.

Maha Salem, Gabriella Lakatos, Farshird Amirabdollahian, and Kerstin Dautenhahn. 2015. Would You Trust a (Faulty) Robot?: Effects of Error, Task Type and Personality on Human-Robot Cooperation and Trust. In *Proceedings of the 10th Annual ACM/IEEE International Conference on Human-Robot Interaction (CHII)*. ACM, Portland (OR), USA, 141–148.

Mark Scheeff, John Pinto, Kris Rahardja, Scott Snibbe, and Robert Tow. 2002. Experiences with Sparky, a Social Robot. In *Socially Intelligent Agents - Creating Relationships with Computers and Robots,* Kerstin Dautenhahn, A. Bond, L. Cañamero, and B. Edmonds (Eds.). Springer, Boston (MA), USA, Chapter 21, 173–180.

Sarah Sebo, Margaret Traeger, Malte Jung, and Brian Scassellati. 2018. The Ripple Effects of Vulnerability: The Effects of a Robot's Vulnerable Behavior on Trust in Human-Robot Teams. In *Proceedings of the 13th ACM/IEEE International Conference on Human-Robot Interaction*. ACM, Chicago (IL), USA, 178–186.

Amanda Sharkey and Noel Sharkey. 2020. We need to talk about deception in social robotics! *Ethics and Information Technology* (2020), 1–8. https://doi.org/10.1007/s10676-020-09573-9

Rosanne M Siino, Justin Chung, and Pamela J Hinds. 2008. Colleague vs. tool: Effects of disclosure in human-robot collaboration. In *The 17th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN).* IEEE, Munich, Germany, 558–562.

Matthew Noah Smith. 2010. Reliance. *Noûs* 44, 1 (2 2010), 135–157.

Herman Tavani. 2018. Can Social Robots Qualify for Moral Consideration? Reframing the Question about Robot Rights. *Information* 9, 4 (3 2018), 73.

Suzanne Tolmeijer, Astrid Weiss, Marc Hanheide, Felix Lindner, Thomas M Powers, Clare Dixon, and Myrthe L Tielman. 2020. Taxonomy of Trust-Relevant Failures and Mitigation Strategies. In *15th ACM/IEEE International Conference on Human-Robot Interaction (HRI).* ACM, Cambridge, United Kingdom, 3–12.

Margaret L Traeger, Sarah Strohkorb Sebo, Malte Jung, Brian Scassellati, and Nicholas A Christakis. 2020. Vulnerable robots positively shape human conversational dynamics in a human-robot team. *Proceedings of the National Academy of Sciences of the United States of America* 117, 12 (3 2020), 6370–6375.

Sherry Turkle. 2011. *Alone Together: Why We Expect More From Technology and Less From Each Other.* Basic Books, New York (NY), USA. 400 pages.

Samuele Vinanzi, Massimiliano Patacchiola, Antonio Chella, and Angelo Cangelosi. 2019. Would a robot trust you? Developmental robotics model of trust and theory of mind. *Philosophical Transactions of the Royal Society B* 374, 1771 (2019), 1–9.

Blay Whitby. 2008. Sometimes it's hard to be a robot: A call for action on the ethics of abusing artificial agents. *Interacting with Computers* 20, 3 (2008), 326–333.

Jakub Złotowski, Hidenobu Sumioka, Shuichi Nishio, Dylan F Glas, Christoph Bartneck, and Hiroshi Ishiguro. 2016. Appearance of a robot affects the impact of its behaviour on perceived trustworthiness and empathy. *Paladyn* 7, 1 (2016), 55–66.

Challenges and solutions for trustworthy explainable robots

Guglielmo Papagni, Sabine T. Koeszegi

Abstract

For robots to be accepted within society, non-expert users must deem them not only useful (and usable) but also trustworthy. Designing robots that can explain their decisions and actions in terms that everyone can understand is crucial to their trustworthiness and successful integration into our society. This paper, written as a part of a doctoral dissertation, draws from interdisciplinary research on social sciences and explainable robots (and AI) to address the set of challenges associated with making robots explainable and trustworthy. Particular attention is paid to non-expert users' perspectives within the context of everyday interactions. We claim that, as perfect explanations do not exist, their success in triggering understanding and fostering trust is determined by their plausibility. Furthermore, we maintain that plausible explanations are the result of contextual negotiations between the parties involved. As a result, this paper presents strategies formalized into a model for explanatory interactions to maximize users' understanding and support trust development.

Keywords

Explainable Robots, Trust, Non-Expert Users, Everyday Explanations

1 Introduction

Recently, the concept that AI and robots should be able to explain their inner workings, decisions, and actions has emerged in academic and societal discussions. Furthermore, as AI and robots permeate society at different levels, affecting people's everyday life, their decision-making processes should be understandable not only for machine learning and robotics experts but also for a broader audience of domain experts (i.e., practitioners from fields where AI technologies are applied) and non-expert end-users. Importantly, each of these categories of users has different demands in terms of explainability desiderata and goals, as their interests and knowledge of the technology may differ substantially. To this extent, it is crucial to understand and acknowledge the differences between different categories of users and, hence, what explainability entails in each context.

The category of domain experts is concerned with applications, such as military operations (e.g., robots used for mine detection and removal or rescue tasks), exploration (e.g., in space or the oceans), and medical purposes. This implies that most of the users will need to undergo some sort of special training to interact with the robots. While this does not guarantee that these users will become robotics experts, such a training allows for creating an adequate mental model of the robot that, in turn, supports users' understanding and trust calibration. In contrast, the category of non-expert users refers to users who have little to no previous experience with specific robotic technologies. It includes application contexts such as caregiving and education, recreational activities, and, perhaps

https://doi.org/10.34727/2022/isbn.978-3-85448-052-5_3

most crucially, interactions with robots 'in the wild' [Sabanovic et al. 2006]. Because there has been no previous interaction or any introduction, the level of uncertainty concerning robots is higher in these contexts. According to several definitions, uncertainties and perception of risk represent two elements that may jeopardize trust [Lee and See 2004; Andras et al. 2018; Luhmann 2018].

This paper addresses a set of challenges of making robots explainable and trustworthy, particularly for non-expert users and within the context of everyday interactions. The main reason for doing this project is those non-expert users represent the vast majority of the public, and many robots and other AI-based technologies are designed to interact with them daily. Furthermore, because of their lack of technical knowledge and agency to manipulate robotic technologies, non-expert users are the most vulnerable. In this context, explainability plays a crucial and multifaceted role. According to some studies, explanations that are properly tailored to the needs of non-expert users reduce perceived uncertainty and increase the understandability of robots. This, in turn, supports users with trust calibration toward robots and, consequently, robot acceptability in society [Lomas et al. 2012; Langley 2016; Langley et al. 2017; Sheh 2017b; Andras et al. 2018; Papagni and Koeszegi 2020, 2021b]. Therefore, designing robots that can explain their decisions and actions in terms that everyone can understand will aid in their successful integration into our society. Furthermore, while the interests and needs of specific groups of users might differ, an explanation that is understandable by users with no prior knowledge of robotic technologies should be understandable to more technologically accustomed ones.

One of the major problems in tailoring robot explanations to the needs of non-expert users is that explainability is frequently considered a data-driven rather than goal-driven characteristic [Sado et al. 2020]. Instead, we claim that the design of social robots should integrate inputs from various disciplines and focus on developing the capacity to communicate decisions in terms easily graspable by a broad audience. Another problem that requires more extensive investigation is that explanations are, by their very nature, incomplete approximations of the actual decision-making processes [Keil 2006; Rudin 2018; Wang 2019]. The lack of perfect explanations is even more problematic for robotics, given the standardized, algorithmic, and 'coordinate-based' modalities of information processing that are typical of robots [Lomas et al. 2012].

We approach these challenges with an interdisciplinary drive. Seeking and providing explanations is a form of everyday social communication, which has been extensively studied within disciplines, such as philosophy, sociology, and psychology [Hilton 1990; Miller 2019]. Combining findings from such disciplines with the need to integrate them into the design of robots and other artificial agents can be labeled as an 'interdisciplinary challenge' of explainability [Adadi and Ber-

rada 2018]. Specifically, this paper discusses the core elements of a recent model for explanatory interactions with artificial agents proposed by the authors of this paper (see figure 1). The remainder of this paper is organized as follows. Section 2 introduces the model and briefly analyzes its development and core elements. Concerning the standardization of explanations, Section 2 presents the concept of contextual, co-constructed plausibility as the most significant feature upon which explanations should be built. Section 3 addresses the timing of explanations, which represents a central element of the model, to answer the question of when explanations are mostly needed to support the trust calibration between users and robots. Furthermore, Section 3 briefly presents the results of a study conducted in the context of repeated interaction with a virtual agent, whose accuracy and explainability are manipulated. Section 4 discusses whether a robot's decision or action ought to be explained because of intentions and reasoning or other causes (e.g., natural or mechanical), as this aspect is critical for the structure of an explanation. Section 5 focuses on communication strategies to increase the explanations' understandability, particularly on the possibility of multimodal and interactive explanations, which is at the heart of the non-expert users' question. Section 6 concludes the paper by addressing limitations and outlining the direction for future work.

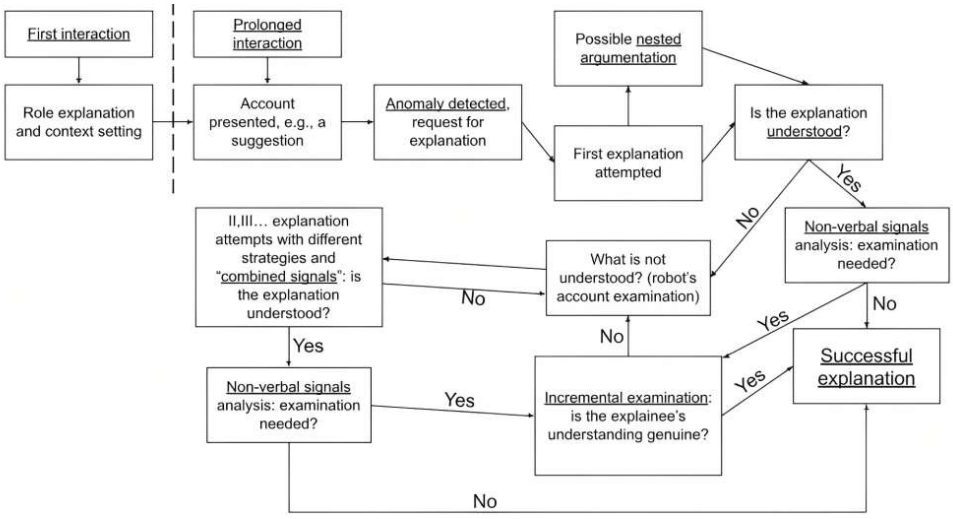


Figure 1 Explanatory Dialogue Model Adapted from [Papagni and Koeszegi 2021b]

2 Explanatory dialogue models

According to Berland, "literature in both the philosophy of science and psychology suggests that no single definition of explanation can account for the range of information that can satisfy a request for an explanation" [Berland and Reiser 2009, p. 27]. Accordingly, there is no single model to describe a perfect explanatory interaction. Furthermore, as previously stated, such models must be suitable for implementation in the algorithmic information-processing units of robots. The model presented in this paper aims to cope with this issue [Papagni and Koeszegi 2021b]. To do so, we analyzed existing models for explainable artificial intelligence (XAI), identified shortcomings, and developed solutions accordingly [Walton 2011; Madumal et al. 2018, 2019].

We identify two major limitations in Walton's, as well as Madumal's, Miller's, Sonenberg's, and Vetere's models. Sections 3 and 5 discuss more thoroughly each of these shortcomings. However, it is important to introduce them, as they both play central roles in the design and structure of our model. The first one concerns the timing of explanatory interactions and, more specifically, the notion that explanation requests are always promoted by an 'anomaly detection' [Walton 2011] or 'knowledge discrepancy' [Madumal et al. 2018, 2019]. This approach expresses the idea of explanations as isolated events, rather than as contextual instances. For instance, the models mentioned do not account for the fact that explanations concerning the robot's function in the specific interaction context are required at the beginning of an interaction with a robot, especially if this occurs 'in the wild'. This moment plays a role in how people build their mental model of the robot and should thus be considered part of the explanatory interaction.

The second shortcoming we identify is ensuring users' understanding of explanations. As previously noted, the inner workings of AI-based technologies are difficult to understand, even for expert users, let alone non-expert. If the robots' explanations are also not properly understood, the initial problem remains, since customers will still be unable to make sense of the robots' behavior. This argument also holds when applied to wrong explanations. How could an explanation be labeled as wrong if the content is not understood? Section 5 addresses these considerations more in detail.

2.1. Explanations' plausibility

Our model leverages on the principle of explanations' plausibility as the key criteria and ultimate goal [Papagni and Koeszegi 2020]. According to Karl Weick's 'sensemaking theory', sensemaking intended as a process, "is driven by plausibility rather than accuracy" [Weick et al. 2005, p. 415]. Building upon Peirce's work

on abductive reasoning, Wilkenfeld and Lombrozo rework Harman's concept of 'inference for the best explanation' [Harman 1965; Peirce 1997; Wilkenfeld and Lombrozo 2015]. Specifically, they postulate that the purpose of explainability should be to provide the best understanding of the causes of an event, rather than the most accurate explanation possible. This approach is consistent with Weick's idea that, to grasp the causes of an event, people seek plausible stories (i.e., that something 'might be') more than they seek true stories (i.e., that something 'actually is') [Peirce 1997; Miller 2019].

Malle argues that people seek explanations to find meanings and manage social interactions [Malle 2006]. According to Weick, the process of building meanings is the result of a collaborative effort involving the two parties (i.e., the explainer and explainee), as well as the context within which the interaction occurs [Weick et al. 2005]. In terms of explanatory interactions, there must be a knowledge transfer from the robotic explainer, who initially and 'asymmetrically' possesses the information that makes a specific explanation plausible, to the explainee, who must understand and agree that the explanation is plausible in that given context and for a specific event [Malle et al. 2007]. This does not necessarily imply that the explanation provided is the best in absolute terms, let alone the only one. The emphasis on all parties involved agreeing on the plausibility of an explanation implies the explainee's understanding of the explanation (i.e., it is unlikely for someone to find something plausible without understanding it in the first place). Furthermore, viewing plausibility as a collaborative and contextual achievement implies that the parties involved judge a given explanation as successful if it provides a satisfying account of an event's most likely causes.

Another advantage of adopting plausibility and abductive reasoning as core criteria of explainability is that there is no universally accepted principle for selecting a subset of causes upon which explanations are built. While certain qualities, such as internal coherence of an explanation and coherence of an explanation with prior beliefs, are generally considered desirable [Thagard 1989; Lombrozo 2007], the choice of other features is less obvious. For instance, some studies emphasize that explanations should be simple, whereas others consider complexity as the trademark of quality [Lombrozo 2007; Kulesza et al. 2013; Zemla et al. 2017]. If an explanation is only considered plausible when all the concerned parties agree, it follows that the most significant qualitative requirements for that situation are met. For an explanation to be (co-)considered plausible, the amount of information it conveys cannot be overwhelming or too scarce. Likewise, the explanation must be coherent with itself and with the prior beliefs of the concerned parties; it must not be too generic and vague, or complex, and so on. However, it could still be that an explanation will not be immediately considered plausible by all concerned parties. As plausibility is a quality that results from a negotiation,

multiple utterances may be required before all parties are satisfied. Section 5 discusses how this limitation can, at least in part, be dealt with.

2.2. Explainable robots, plausible robots

From these last considerations, plausibility is not a property that can be predefined once and for all. In other words, it is an aspect that is mostly determined by the context in which an interaction unfolds, the actors involved, and their specific interests. For instance, a possible application for social robots is assisting library customers. Among other tasks, such robots may suggest new readings to the customers, who may want to know the reasons for a specific recommendation before deciding. In a similar case, if the timing is not an issue, the robot may explain in detail how it arrived at that recommendation, by demonstrating how features, such as the customer's record of books requested in the past or feedback and reviews left by other users with similar preferences, weighed in the decision-making process [Ramos-Garijo et al. 2003; Mikawa et al. 2009; Sreejith et al. 2015]. Once these criteria have been presented by the robot, the customer may eventually agree (or disagree) with the explanation's plausibility and act accordingly.

However, in different situations, other features would likely be more relevant to show an explanation's plausibility. For instance, when the timing is an issue (e.g., during a rescue operation [Murphy 2004]), people may want robots to provide simple and concise explanations while not sparing vital information, particularly if the consequences of a wrong decision are potentially disastrous. In conclusion, what plausibility entails cannot (and probably should not) fall under an unambiguous, umbrella definition. The reason for this is that whether an explanation is plausible or not should be negotiated between the concerned parties, in a specific context.

3 Explanations' timing

This section focuses on the timing of explanations, a critical aspect that previous models have ignored, at least partially. Both models identify the start of an explanatory interaction in a 'knowledge discrepancy' or 'anomaly detection' [Walton 2011; Madumal et al. 2019]. Even though these models envision back-and-forth explanatory interactions with artificial agents, the type of approach they symbolize is one that ideally regards explanations as isolated instances. In contrast, we support Weick's view that meanings are co-constructed in the interplay between the concerned actors and the context, as we explain in the following paragraphs.

3.1. Initial explanations and trust formation

People provide explanations according to their mental model of the person with whom they are interacting in terms of the level of expertise and 'technicality' [Cawsey 1993]. In principle, this process is reliable because the parties involved in an explanatory interaction often share some knowledge about the topic being discussed. However, when it comes to robots, this can be problematic. When robots are employed in semi-controlled environments (e.g., in elderly care facilities or educational contexts), the researchers involved introduce them to users. To help users become acquainted with the robots, the researchers explain what the robots can and cannot do and support users in establishing an adequate initial mental model of the robots.

However, social robots are ultimately supposed to operate also 'in the wild' in everyday situations (e.g., at shopping malls and libraries) where people will mostly have little to no experience with robots and interactions will be limited in time. To this extent, initial trust depends on both personal attitude toward technology and 'institutional cues' [Siau and Wang 2018; Andras et al. 2018]. The former is a consequence of the combination of several factors, such as cultural background, demographics, and personality traits [Morris and Venkatesh 2000; Chien et al. 2016], and it can result in an equally wide range of dispositions toward new technologies, which are not necessarily mediated by accumulated experience with such technologies. These range from high expectations and over-trust [Dzindolet et al. 2003; De Visser et al. 2020], to skepticism and even forms of 'technophobia' [Kerschner and Ehlers 2016].

The notion that trust partially depends on 'institutional cues' refers to the role played by 'third parties', such as private companies, developers working for them, national and international institutions, and experts and regulatory bodies. Leveraging on their reliability and reputation, such entities play a 'proxy' role in determining how people perceive and trust new technologies. Specifically, this process is based on the assumption that the entities introducing new technologies act in accordance with values, such as integrity and benevolence, that define moral trust [Elia 2009; Lankton et al. 2015; Sood 2018]. Researchers have expressed concerns about the transparency, responsibility, and accountability of such 'third parties'. As for end-users initial trust in robots and AI, it is crucial to emphasize the importance of the adequate distribution of responsibilities (to, e.g., ensure technology transparency) among the stakeholders [Elia 2009; O'Leary 2019].

Based primarily on 'institutional cues' and individual attitude, initial trust can be very high or low irrespective of robots' actual performance concerning their purposes (i.e., not calibrated). For this reason, we emphasize the importance of the initial explanations. When robots have not yet proved to be reliable and

benevolent (e.g., on behalf of their makers), initial explanations may substitute the missing previous interactions, support the establishment of adequate mental models, and guide users toward placing calibrated trust in robots [Andras et al. 2018; Fossa 2019].

We agree with Cawsey that, in the event of a first-time interaction, robots should treat users 'as novices' which implies that robots should not assume anything about what users know. Accordingly, the robots' mental models of the users should only evolve and update as an interaction develops [Cawsey 1993]. According to Weick's argument, by adopting this approach, meanings and knowledge are lifted from the private and implicit sphere and made public and explicit [Weick et al. 2005]. Interestingly, Walton notes that "to grasp the anomaly, you have to be aware of the common knowledge" [Walton 2011, p. 365] and that "the system has to know what the user knows, to fill in the gaps" [Walton 2011, p. 365]. This appears to contradict the idea that explanation requests are triggered by the detection of an anomaly in one's account. However, how could a robot know what the user knows? Likewise, how can a user detect an anomaly in a robot's behavior if the user has no prior knowledge of what the robot should or should not do? For this reason, our model proposes that robots should provide initial explanations that contain basic information, such as what role and purpose the robot have and what it can and cannot do (see top left part of Figure 1). By so doing, robots could proactively establish the interaction context and support users in developing an adequate mental model. Additionally, once users are informed, the basic notions about the robot become shared knowledge and the robot can update its mental model of the user accordingly.

3.2. Unexpected events and trust restoration

According to the literature, the other moment in an interaction when people seek out explanations is when something unexpected or unpredictable happens [Andras et al. 2018; Miller 2019]. In other words, once users establish a mental model of a robot based on prior interactions, they will expect the robot to perform actions within a certain range of possibilities. Within this range, the robot's reliability will be progressively determined based on its performance and accuracy. As a robot regularly demonstrates reliability and trustworthiness, users may consolidate their positive mental model of it, so that explanations become superfluous if not even damaging [Doshi-Velez and Kim 2017]. However, a robot may still act unexpectedly or unpredictably. Such events, which do not fit into the established mental model, are also recorded in the interactions. This is also what the models by [Walton 2011; Madumal et al. 2018, 2019] label as 'knowledge discrepancy' or 'anomaly'. In similar situations, users' understanding of the robot's behavior

is challenged. Furthermore, several researchers have found that if users do not understand why a robot is behaving in a certain way, their acceptance and trust in the robot are probably weakened [Lomas et al. 2012; De Graaf and Malle 2017; de Graaf et al. 2018; Miller 2019]. While this is particularly the case when unexpected robot actions turn out to be mistakes [Elangovan et al. 2007; Robinette et al. 2017] even if a robot behaves according to its internal planning, if this is not obvious to users, it is important that they still make sense of why the robot is acting that way [Andras et al. 2018].

In an aging society, social robots are meant to be deployed in elderly-care facilities with assisting duties. IBM's MERA is one such robot being developed, on top of SoftBank Robotics' Pepper platform, for similar purposes. It can monitor people's pulse and breathing functions, among other things [Martinez-Martin and del Pobil 2018; Venkatesh 2019]. For instance, if the robot detects any anomalies in these parameters before the person is consciously aware of it, it may suggest the assisted person rest. Such an event may be perceived as an anomaly, prompting the assisted person to request an explanation, which would likely elucidate the reasons behind the suggestion and show that, while these reasons were not obvious at a first glance, they still make the robot's suggestion plausible.

Hence, whether it is to prevent the loss of trust, or restore it after a mistake, robots must provide reasons for their actions through explanations. Other trust restoration strategies, such as denial, apologies, compensation, and relationship restructuring, exist and can be implemented among robots' functions [Quinn et al. 2017; Lewicki and Brinsfield 2017]. However, unlike these strategies, explainability offers two main advantages. On the one hand, as we discussed in the previous paragraph, explanations support trust not only in the case of a violation but also in building it at the start of an interaction. On the other hand, explanations provide useful insights into the causes of an unexpected event or mistake. We previously noted that explanations may not be strictly necessary in case of repeated successful interactions with a robot. For instance, when "there are no significant consequences for unacceptable results" [Doshi-Velez and Kim 2017, p.3] or when a problem has been thoroughly researched and validated in real-world scenarios, explanations could become superfluous. However, even after multiple interactions, specific users may be unaware that a certain problem has been previously studied and that a robot's decision is based on real-world-validated data. Therefore, in principle, robots should always be able to explain themselves whenever users ask.

To examine some claims discussed in the previous sections, an empirical study was conducted. Participants were required to interact with a personalized virtual learning assistant seven times. The goal of the assistant was to provide participants with recommendations on what chunks of text to focus on (out of

larger portions), for them to prepare for quizzes. The system's explainability and accuracy were modified throughout the study.

Among the main findings, we observed that, contrary to expectations, initial explanations about the system's functionality did not increase initial trust. Simultaneously, the assistant's wrong recommendation affected participants' trust negatively, as it was perceived as a trust breach. However, qualitative data reveal that participants tended to be quite tolerant toward imperfect AI-based systems, as these systems are not expected to always function perfectly. Additionally, the qualitative data suggest that the researchers' 'hidden authority' has a favorable impact on the system's trustworthiness. Perhaps more importantly, trust restoration was significantly faster when the system provided an explanation following the wrong recommendation, rather than not. Specifically, explanations were the most effective as a trust-restoration strategy with risk-averse participants. Furthermore, explanations aided trust recovery, even if the participants did not always access them. Our qualitative analysis revealed how this may be explained, at least in part, by the fact that the very availability of explanations suggests a more transparent and trustworthy system.

4 Explainable robots and the intentional framework

Another element is crucial in terms of the mental model of robots and explanation generation. It is about whether or not robots' explanations ought to reflect some form of intentionality (and other mental states) behind robots' behavior. This aspect of explanatory interactions is part of a broader ongoing discussion between the human-robot and human-computer interaction (respectively, HRI and HCI) communities. While discussions on robots' 'mental states' have paced up recently, they have older roots that date back at least to Heider's and Simmel's work, as they demonstrated that people adopt a mentalistic framework to interpret even the movements of simple and schematic geometrical shapes [Heider and Simmel 1944]. Then, with Daniel Dennett's concept of the 'intentional stance', interest in the topic has spread. [Dennett 1988, 1989]. Dennett explained that people interact with certain technological artifacts (such as a chess-playing computer) as though they acted on human-like internal states, such as desires, beliefs, and intentions. According to Dennett, it would be too difficult to understand how such devices work solely by relying on one's knowledge of their intended purpose (i.e., the design stance), let alone the knowledge of natural laws (i.e., the physical stance) that ultimately govern everything [Dennett 1988, 1989, 1997]. Therefore, Dennett says, people adopt with computers and robots a mentalistic framework that is similar to that adopted with other people.

According to recent interpretations, the phenomenon is due to a 'primacy of the social mindset', which means that a mentalistic interpretative framework is always readily available because of people's social training and familiarity with it since childhood [Buckner et al. 2008; Looser and Wheatley 2010; Spunt et al. 2015; Papagni and Koeszegi 2021a]. Furthermore, as most people appear to lack a strategy for interacting specifically with sophisticated technologies, such as robots, a mentalistic approach eventually prevails. Attributing intentions to robots and other seemingly intelligent machines has some problematic aspects. For instance, researchers have proposed that in certain cases, the unconscious (and erroneous) adoption of a mentalistic framework may be the origin of the so-called 'uncanny valley' phenomenon [Bartneck et al. 2009; Mori et al. 2012]. Additionally, in certain situations, attempting to understand robots' behavior from a mentalistic perspective is not the best strategy, and users may have to forcefully adapt their mental model at the expense of cognitive resources [Wiese et al. 2017].

According to Weick's sensemaking framework, finding meanings in the social context of everyday life entails bringing order to the chaotic stream of both intentional behaviors and unintentional events. In terms of explanations, this translates to attributing either reasons, intentions, desires, and beliefs, or natural and mechanical causes. According to De Graaf and Malle, intentionality is a core concept that allows people to explain and understand others' behaviors [De Graaf and Malle 2017]. While the phenomenon has been thoroughly investigated in the human sciences, the concept of predicting and explaining robots' behavior using the intentionality framework is an open debate. According to Bossi, "people may treat robots as mechanistic artifacts or may consider them to be intentional agents. This might result in explaining robots' behavior as stemming from operations of the mind (intentional interpretation) or as a result of mechanistic design (mechanistic interpretation)" [Bossi et al. 2020, p. 1].

As we previously discussed, explanations are often sought after when users' mental models of robots are challenged by unpredictable events. This includes situations in which users cannot understand or explain robots' actions according to the mental model of robots they already possess. An implication of this interpretative gap is that whatever framework (i.e., intentional or mechanistic) users are adopting at the time of the unexpected occurrence, their trust in the framework's prediction-making power might decrease. In other words, when something unexpected happens, users may be unable to provide themselves with reasons or causes and, hence, ask the robot with whom they are interacting for an explanation. Some cases will force a complete perspective (i.e., framework) switch, while others will not. Importantly, according to De Graaf and Malle, robots "must be able to distinguish intentional from unintentional behaviors" and they "must be able to explain each of these classes of behavior in the expected way – unin-

tentional behaviors with (mere) causes, intentional behaviors with reasons" [De Graaf and Malle 2017, p. 19].

For instance, We previously mentioned, referring to elderly people's assistance, the possibility of the robot advising the assisted person takes a rest. The latter may not immediately grasp the reason for the recommendation, as they are unaware of what the robot knows. This includes not knowing whether the recommendation is genuine (i.e., based on the intention to assist the person) or based on a wrong premise (e.g., a malfunction). Assuming that the robot has been useful and has acted in the best interest of the user up to that point, the user may be struggling to make sense of the recommendation within the same (i.e., intentional) framework and may request an explanation. Within an intentional framework, the robot's explanation that its sensors have observed increased heart rate and heavy breathing would still make sense, as it would show the robot's intention to assist the user. A similar explanation emphasizes that the user was merely unaware of the robot's actual decision-making process. Accordingly, this implies that not every unpredictable behavior is the result of robots' malfunctions or internal errors, which are more likely to be detected (e.g., if the robot suddenly stops performing its tasks), and require users' to switch framework.

Ultimately, it could still be that a robot provides an explanation that makes sense (i.e., sounds plausible) within the boundaries of the framework adopted by the users but is built upon wrong premises [Dunne et al. 2005; Walton 2011]. As will be discussed in Section 5, when dealing with the structure of explanatory interactions, the risk of wrong explanations going unnoticed motivates taking further measures. Based on the discussion in this section, we claim that robots must be designed to support users, by means of explanations, in adopting the most appropriate interaction framework. This is especially the case for the early stages of extensive adoption of robotics in everyday contexts. Indeed, these times are most characterized by uncertainty in terms of both the adoption of and narratives built around these technologies. Furthermore, whenever necessary, robots should support the transition from one interpretative framework to another. We have previously discussed how the plausibility of explanations must be considered a contextual joint achievement. What framework is most adequate for understanding an event is a contextual feature that must be treated as such. Hence, robots should communicate explicitly and clearly, to the greatest extent feasible, whether the event being explained involves unintentional causes (e.g., an internal failure or mistake or uncontrollable external forces) or intentional reasons. In the next section, we will discuss explanation communication strategies that maximize the chances of users' correct understanding and hence trust toward the robots.

5 Communicating explanations

Explanations are primarily forms of social communication [Hilton 1990]. Therefore, addressing how robots should deliver explanations is likely the most essential aspect of explanatory interactions. This section analyzes two features of explanation communication that constitute the core of our model. Specifically, we discuss our claims that to support users' understanding and trust calibration, robots should:

- Be able to use diversified means of communication.
- Provide users with the possibility to question explanations and ask for further insights.

Importantly, when it comes to explainable robots, the research on the effects of combining the two mentioned strategies while promising is still in its early stages [Abdul et al. 2018; Anjomshoae et al. 2019].

5.1. Multi-modal explanation

In human-human interactions, explanations are primarily communicated through natural language. Generally, they should follow communication norms, such as 'Grice's (four) maxims of conversation' [Grice 1975]. They refer to communicating only what is confidently believed to be accurate, avoiding overwhelming amounts of information without being scarce, relevant to the context (i.e., a 'good social explanation' [Hellström and Bensch 2018; Miller 2019]), avoiding obscurity and ambiguity and being brief and orderly in presenting the information. Grice's maxims are often mentioned in explainable robots and AI research because they provide an implementable solution that may improve explanation quality [Miller 2019; Papagni and Koeszegi 2021b]. Sheh provides further possibilities for modifying how explanations are communicated through natural language [Sheh 2017a]. According to the author, robots can modify the depth and type of explanation based on the needs of specific interaction instances and the availability of the robots' underlying AI models. The author observes, in reference to a scenario in which a robotic shopping mall assistant is questioned about its product recommendations, that in similar circumstances, social robots' explanations are expected to primarily satisfy users' curiosity and support further engagement. For this reason, the author continues, 'Post-Hoc' explanations at 'Attribute Only' or 'Attribute Use' depths may be appropriate for the purpose [Sheh 2017a]. While the former indicates explanations that are tailored solely to what the robot deems the most relevant features, the latter considers the implications (i.e., 'use') of each attribute's value. Therefore, if properly tailored, text-based explanations alone already provide various customization options.

However, when the explanations' goal is to maximize users' understanding and trust calibration toward a robot, it is important to note that natural language only covers a subset of feasible communication strategies. Explanations in the form of 'combined signals' [Engle 1998], also known as 'multi-modal' explanations represent a promising but under-explored research avenue. Anjomshoae, Najjar, Calvaresi, and Främling discussed six possible communication modalities [Anjomshoae et al. 2019]. Besides text-based natural language explanations, they identified the "visualization" (i.e., graphical) type as the second most common one. Logs, expressive motions, expressive lights, and speech complete the list. The notion behind multimodality is that, as technological devices, robots can convey information through complementary modalities, sometimes even better than humans can. For instance, with visual explanations being the second most common after text-based ones, many robots can display on frontal screens graphic information gathered by their sensors, and once processed, these environmental data may support text to convey more complete messages. In our previous example, the IBM's MERA robot explained to the assisted person that its recommendation to take a rest was based on factors, such as the unusually high pulse rate and heavy breathing. While a text-based explanation would likely suffice to convey the essential message, the explanation's quality could still improve if the robot would provide visualizations of the actual scans of normal and abnormal heart activity. While HRI research on multimodality and 'combined signals' is still in its early stages, an increasing number of studies have demonstrated that users can benefit from multimodal explanations. The HCI community has done most of the research in multimodal explanations so far. Most studies effectively combined verbal and visual information, showing how people preferred this format to 'uni-modal' ones [Huk Park et al. 2018; Kanehira et al. 2019].

Two considerations must be made. First, the availability of alternative communication strategies should not mean that robots must display all available information at once. Explanations should not exclude vital information, but simultaneously, they should also not overwhelm users with too much information. To this extent, researchers propose that, in certain cases, employing alternative single-handed modalities may be more beneficial to the users. For instance, referring to robots' reactive planning, Theodoru, Wortham, and Bryson suggest that since robots can take many decisions per second, graphical explanations are more efficient and direct than verbal ones [Theodorou et al. 2016]. Giving self-driving systems the ability to employ light signaling to communicate simple messages to pedestrians, such as that they can cross the street safely [Faas and Baumann 2019], is another example of how alternative modalities can suffice even when taken alone. In conclusion, while in certain cases alternative modalities may provide adequate information, text-based explanations are likely to remain prominent (possibly sup-

ported by other means) because the semantic richness that can be conveyed through natural language is difficult to match through other means alone.

Finally, multimodality should not be unidirectional or limited to the combination of text-based and graphic communication. Natural language processing and image recognition have improved significantly recently, allowing robots and virtual agents to provide progressively better answers to users' text- or image-based inputs. One further possibility is that robots can 'read' and 'express' signals other than graphic and natural language communications. For instance, research in other relevant areas of robotics, such as (reading and expressing) body motion [Han et al. 2012; McColl and Nejat 2014] or facial expressions and gaze [Fiore et al. 2013; Admoni and Scassellati 2017] shows that robots can process various signal typologies that can make communication with humans (included explanatory interactions) more flexible and inclusive.

5.2. Interactive explanations

Making explanations 'interactive' is another promising strategy to increase robots' explanations quality that requires further investigation, particularly in the field of social robots [Abdul et al. 2018; Papagni and Koeszegi 2021b]. This research is partly driven by the desire to achieve a higher degree of human likeness [Madumal et al. 2018, 2019]. Indeed, explanations in robotics are often treated as 'single-shot' communication acts, whereas in human-human interaction, they frequently occur in the form of dialogues with back-and-forth iterations. However, interactivity also represents a strategy to deal with what Keil identifies as people's attitude to overestimate their own understanding of explanations (i.e., the 'illusion of explanatory depth') [Keil 2003]. According to Keil, this phenomenon, which is related to studies from social psychology on the 'introspection illusion' [Pronin 2009], consists of wrongly assessing the quality of the information one retains after being provided an explanation. The next paragraphs discuss our claim that, among other advantages, im- plementing design features that support interactivity of robots' explanations helps mitigate this phenomenon.

A fundamental contribution to the user-friendliness of explanations' interactivity is that it allows the parties involved to seek further insights to better understand what is being explained, and it allows questioning of both parties' accounts. The implementation of 'nested argumentation dialogues' [Madumal et al. 2018, 2019] and an 'examination phase' into our model aims to primarily tackle this multifaceted aspect [Dunne et al. 2005; Walton 2006, 2011].

Introducing nested argumentation dialogues allows users to engage in multilayered explanations in which they can drift from one question to another in a

back-and-forth manner. This back-and-forth movement may concern the topic of the original question or may be about 'spin-off' discussions [Madumal et al. 2018, 2019]. Often in human-human interaction, such spin-off argumentation dialogues are nested on top of the original explanation to support explainees by improving their understanding. The model proposed by Walton does not account for nested argumentation because the author labels overlapping dialog as an illicit dialectical shift, implying that the previous question must be considered closed [Walton 2011]. However, to achieve interaction naturalness and support users' sensemaking, robots should be able to process nested dialogs as such, leaving users the choice to return to the original one. Hence, to increase the human-likeness of explanatory interactions, our model allows users to engage in nested argumentation dialogues that are both related and unrelated to the original question, as shown in the top right corner of Figure 1. However, introducing such internal loops is merely one interpretation of the concept of interactivity.

Explanations may appear logical at a first glance and yet be grounded upon incorrect premises [Walton 2011; Dunne et al. 2005; Lakkaraju and Bastani 2020]. Introducing a dialectical shift in the form of an 'examination phase' allows users to analyze the explainer's account for any inconsistencies and evaluate the quality of the explanation for potential errors [Dunne et al. 2005; Lamche et al. 2014]. To this extent, Kaur et al. highlight the propensity, even among HCI expert practitioners, to over-rely on interpretability tools' visual outputs in a study in which they analyze participants' reactions to different approaches to model interpretability (i.e., 'glass-box' and 'black-box') To address this issue, one of their suggestions is to adopt 'back-and-forth explanations' (i.e., interactive interpretability) [Kaur et al. 2020].

Another possible use for an examination phase is to test the explainee's understanding of an explanation, as suggested by Walton [Walton 2011]. Indeed, as previously stated, people are susceptible to the 'illusion of explanatory depth' and tend to overestimate their understanding of explanations [Keil 2003]. Section 1 also highlighted the connections between understanding robots (and robots' explanations) and calibrating trust in them. For these reasons, assessments of understanding quality are an important aspect of models for explanatory interactions. This is supposed to be done by questioning the explainee about the explanation, the causal connections to the event being explained, and so on. Nevertheless, testing users' understanding should not translate into an interrogation, as this may be perceived as aggressive and have overall counterproductive effects on the interaction [Walton 2011]. To this extent, the authors of the model described in [Madumal et al. 2018, 2019] assert that such an operation is uncommon in everyday human-human interactions. Instead, to keep the interaction as natural as possible, they consider the explainee's affirmation of effective under-

standing as a sufficient criterion to measure the quality of the explanation. While we agree that explanatory interactions should feel natural and smooth to users, rather than making them feel uncomfortable and jeopardizing future interactions, we also acknowledge a gap in the model from Madumal, Miller, Sonenberg, and Vetere in terms of evaluation strategies for the success of an explanation. Therefore, we deem an 'incremental approach' to be the most appropriate [Papagni and Koeszegi 2021b]. Alternatively, after a robot provides an explanation, it may ask users to pick among multiple options what they understood to be the right explanation. To this end, we claim that testing users' understanding must be contextually calibrated based on how much time and interest users are willing to invest. In other words, instead of being predetermined by the robot, questions concerning the explanation must be negotiated with users based on contextual affordances.

Finally, just as it occurs in human-human interaction, it is impossible to guarantee the success of explanations in terms of knowledge transfer and users' understanding. Despite robots' best attempts, there will be circumstances in which users do not grasp what is being explained to them. Future research on explainable robots should focus on how to minimize the likelihood of such events occurring by refining and testing solutions, such as the ones presented in this paper, and implementing alternative strategies to better prevent trust losses and restore trust after a violation.

6 Future work and conclusions

The presence of social robots in everyday life is becoming a reality. Their successful integration and acceptability into society depend not only on how useful they prove to be in terms of performance but also on how they explain their decisions to a broad audience of non-expert users. At the same time, this paper acknowledges that perfect explanations do not exist and that making robots explainable poses a multifaceted interdisciplinary challenge. To solve this problem, we proposed a model for explanatory interactions. This model considers important findings from social sciences as well as from research on explainable AI and robots and their affordances and availability in terms of explainability. Furthermore, as the key criterion to assess the quality of explanations, we proposed a notion of explanations' plausibility as a joint achievement, which presupposes the users' understanding of robots' explanations.

One of the main limitations is that the type of explanation a robot can provide depends on the availability of the underlying algorithms and the physical capabilities of individual robots. In other words, not all the features of our model may

be implemented in the behavioral programming of certain robots. Therefore, research should focus on how to broaden the scope of both AI models' explainability and robots' customization. Another limitation concerns the primarily conceptual nature of the work presented in this paper. This calls for follow-up experimental studies to test our claims and the feasibility of implementing the various features of our model. Such studies shall, for instance, focus on the long-term effects of explanations on trust formation and restoration. Likewise, the combination of multimodal and interactive strategies is a promising but understudied research avenue that may shed further light on users' reception of explainable robots in terms of both trust and understandability.

Bibliography

Ashraf Abdul, Jo Vermeulen, Danding Wang, Brian Y Lim, and Mohan Kankanhalli. 2018. Trends and trajectories for explainable, accountable and intelligible systems: An hci research agenda. In *Proceedings of the 2018 CHI conference on human factors in computing systems.* 1–18.

Amina Adadi and Mohammed Berrada. 2018. Peeking Inside the Black-Box: A Survey on Explain- able Artificial Intelligence (XAI). *IEEE Access* 6 (2018), 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

Henny Admoni and Brian Scassellati. 2017. Social eye gaze in human-robot interaction: a review. *Journal of Human-Robot Interaction* 6, 1 (2017), 25–63.

Peter Andras, Lukas Esterle, Michael Guckert, The Anh Han, Peter R. Lewis, Kristina Milanovic, Terry Payne, Cedric Perret, Jeremy Pitt, Simon T. Powers, Neil Urquhart, and Simon Wells. 2018. Trusting Intelligent Machines: Deepening Trust Within Socio-Technical Systems. *IEEE Technology and Society Magazine* 37, 4 (2018), 76–83. https://doi.org/10.1109/MTS.2018.2876107

Sule Anjomshoae, Amro Najjar, Davide Calvaresi, and Kary Främling. 2019. *Explainable Agents and Robots: Results from a Systematic Literature Review*. International Foundation for Autonomous Agents and Multiagent Systems. http://dl.acm.org/citation.cfm?id=3306127.3331806

Christoph Bartneck, Takayuki Kanda, Hiroshi Ishiguro, and Norihiro Hagita. 2009. My robotic doppelgänger-A critical look at the uncanny valley. In *RO-MAN 2009-The 18th IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, 269–276.

Leema Kuhn Berland and Brian J Reiser. 2009. Making sense of argumentation and explanation. *Science education* 93, 1 (2009), 26–55.

Francesco Bossi, Cesco Willemse, Jacopo Cavazza, Serena Marchesi, Vittorio Murino, and Agnieszka Wykowska. 2020. The human brain reveals resting state activity patterns that are predictive of biases in attitudes toward robots. *Science robotics* 5, 46 (2020).

Randy L Buckner, Jessica R Andrews-Hanna, and Daniel L Schacter. 2008. The brain's default network: anatomy, function, and relevance to disease. *Annals of the New York Academy of Sciences* 1124 (2008), 1–38. https://doi.org/10.1196/annals.1440.011

Alison Cawsey. 1993. User modelling in interactive explanations. *User Modeling and User-Adapted Interaction* 3, 3 (1993), 221–247. https://doi.org/10.1007/BF01257890

Shih-Yi Chien, Katia Sycara, Jyi-Shane Liu, and Asiye Kumru. 2016. Relation between trust attitudes toward automation, Hofstede's cultural dimensions, and big five personality traits. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting,* Vol. 60. SAGE Publications Sage CA: Los Angeles, CA, 841–845.

Maartje MA De Graaf and Bertram F Malle. 2017. How people explain action (and autonomous intelligent systems should too). In *2017 AAAI Fall Symposium Series.*

Maartje MA de Graaf, Bertram F Malle, Anca Dragan, and Tom Ziemke. 2018. Explainable robotic systems. In *Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction.* 387–388.

Ewart J De Visser, Marieke MM Peeters, Malte F Jung, Spencer Kohn, Tyler H Shaw, Richard Pak, and Mark A Neerincx. 2020. Towards a theory of longitudinal trust calibration in human–robot teams. *International journal of social robotics* 12, 2 (2020), 459–478.

Daniel C Dennett. 1988. Précis of the intentional stance. *Behavioral and brain sciences* 11, 3 (1988), 495–505.

Daniel C Dennett. 1989. *The intentional stance.* MIT press.

Daniel C Dennett. 1997. True Believers: The Intentional Strategy and Why It works. *Mind Design* (1997), 57–79.

Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint* arXiv:1702.08608 (2017).

Paul E Dunne, Sylvie Doutre, and Trevor Bench-Capon. 2005. Discovering inconsistency through examination dialogues. In *Proceedings of the 19th international joint conference on Artificial intelligence*. 1680–1681.

Mary T Dzindolet, Scott A Peterson, Regina A Pomranky, Linda G Pierce, and Hall P Beck. 2003. The role of trust in automation reliance. *International journal of human-computer studies* 58, 6 (2003), 697–718.

A R Elangovan, Werner Auer-Rizzi, and Erna Szabo. 2007. Why don't I trust you now? An attributional approach to erosion of trust. *Journal of Managerial Psychology* (2007). 22(1), 4–24. https://doi.org/10.1108/02683940710721910

John Elia. 2009. Transparency rights, technology, and trust. *Ethics and Information Technology* 11, 2 (2009), 145–153.

Randi A Engle. 1998. Not channels but composite signals: Speech, gesture, diagrams and object demonstrations are integrated in multimodal explanations. In *Proceedings of the twentieth annual conference of the cognitive science society.* 321–326.

Stefanie M Faas and Martin Baumann. 2019. Yielding light signal evaluation for self-driving vehicle and pedestrian interaction. In *International Conference on Human Systems Engineering and Design: Future Trends and Applications.* Springer, 189–194.

Stephen M Fiore, Travis J Wiltshire, Emilio JC Lobato, Florian G Jentsch, Wesley H Huang, and Benjamin Axelrod. 2013. Toward understanding social cues and signals in human–robot interaction: effects of robot gaze and proxemic behavior. *Frontiers in psychology* 4 (2013), 859.

Fabio Fossa. 2019. I don't trust you, you faker! On trust, reliance, and artificial agency. *Teoria*, 1(XXXIX) (2019), 63–80.

Herbert P Grice. 1975. Logic and conversation. In *Syntax and semantics. Vol. 3, Speech acts*, P. Cole und J. L. Morgan (Ed.). Brill, 41–58.

JingGuang Han, Nick Campbell, Kristiina Jokinen, and Graham Wilcock. 2012. Investigating the use of non-verbal cues in human-robot interaction with a Nao robot. In *2012 IEEE 3rd International Conference on Cognitive Infocommunications (CogInfoCom).* IEEE, 679–683.

Gilbert H Harman. 1965. The inference to the best explanation. *The philosophical review* 74, 1 (1965), 88–95.

Fritz Heider and Marianne Simmel. 1944. An Experimental Study of Apparent Behavior. *The American Journal of Psychology* 57, 2 (1944), 243–259.

Thomas Hellström and Suna Bensch. 2018. Understandable robots-what, why, and how. *Paladyn, Journal of Behavioral Robotics* 9, 1 (2018), 110–123.

Denis J Hilton. 1990. Conversational processes and causal explanation. *Psychological Bulletin* 107, 1 (1990), 65.

Dong Huk Park, Lisa Anne Hendricks, Zeynep Akata, Anna Rohrbach, Bernt Schiele, Trevor Darrell, and Marcus Rohrbach. 2018. Multimodal explanations: Justifying decisions and pointing to the evidence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 8779–8788.

Atsushi Kanehira, Kentaro Takemoto, Sho Inayoshi, and Tatsuya Harada. 2019. Multimodal explana- tions by predicting counterfactuality in videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 8594–8602.

Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna Wallach, and Jennifer Wortman Vaughan. 2020. Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems.* 1–14.

Frank C Keil. 2003. Folkscience: Coarse interpretations of a complex reality. *Trends in cognitive sciences* 7, 8 (2003), 368–373.

Frank C Keil. 2006. Explanation and understanding. *Annual Review of Psychology.* 57 (2006), 227–254.

Christian Kerschner and Melf-Hinrich Ehlers. 2016. A framework of attitudes towards technology in theory and practice. *Ecological Economics* 126 (2016), 139–151.

Todd Kulesza, Simone Stumpf, Margaret Burnett, Sherry Yang, Irwin Kwan, and Weng-Keen Wong. 2013. Too much, too little, or just right? Ways explanations impact end users' mental models. In *2013 IEEE Symposium on Visual Languages and Human Centric Computing*. IEEE, 3–10.

Himabindu Lakkaraju and Osbert Bastani. 2020. "How do I fool you?" Manipulating User Trust via Misleading Black Box Explanations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. 79–85.

Béatrice Lamche, Ugur Adıgüzel, and Wolfgang Wörndl. 2014. Interactive explanations in mobile shopping recommender systems. In *Joint Workshop on Interfaces and Human Decision Making in Recommender Systems*, Vol. 14.

Pat Langley. 2016. Explainable agency in human-robot interaction. In *AAAI Fall Symposium Series.*

Pat Langley, Ben Meadows, Mohan Sridharan, and Dongkyu Choi. 2017. Explainable agency for intelligent autonomous systems. In *Twenty-Ninth IAAI Conference*.

Nancy K Lankton, D Harrison McKnight, and John Tripp. 2015. Technology, humanness, and trust: Rethinking trust in technology. *Journal of the Association for Information Systems* 16, 10 (2015), 880-918.

John D Lee and Katrina A See. 2004. Trust in automation: Designing for appropriate reliance. *Human factors* 46, 1 (2004), 50–80.

Roy J Lewicki and Chad Brinsfield. 2017. Trust repair. *Annual Review of Organizational Psychology and Organizational Behavior* 4 (2017), 287–313.

Meghann Lomas, Robert Chevalier, Ernest Vincent Cross, Robert Christopher Garrett, John Hoare, and Michael Kopack. 2012. Explaining robot actions. In *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction.* 187–188.

Tania Lombrozo. 2007. Simplicity and probability in causal explanation. *Cognitive psychology* 55, 3 (2007), 232–257.

Christine E Looser and Thalia Wheatley. 2010. The tipping point of animacy: How, when, and where we perceive life in a face. *Psychological science* 21, 12 (2010), 1854–1862.

Niklas Luhmann. 2018. *Trust and power.* John Wiley & Sons.

Prashan Madumal, Tim Miller, Liz Sonenberg, and Frank Vetere. 2019. A grounded interaction protocol for explainable artificial intelligence. *arXiv preprint* arXiv:1903.02409 (2019).

Prashan Madumal, Tim Miller, Frank Vetere, and Liz Sonenberg. 2018. Towards a grounded dialog model for explainable artificial intelligence. *arXiv preprint* arXiv:1806.08055 (2018).

Bertram F Malle. 2006. *How the mind explains behavior: Folk explanations, meaning, and social interaction*. Mit Press.

Bertram F Malle, Joshua M Knobe, and Sarah E Nelson. 2007. Actor-observer asymmetries in explanations of behavior: New answers to an old question. *Journal of personality and social psychology* 93, 4 (2007), 491–514.

Ester Martinez-Martin and Angel P del Pobil. 2018. Personal robot assistants for elderly care: an overview. In *Personal assistants: Emerging computational technologies,* A Costa, V. Julian & P. Novaris (Ed.). (2018), 77–91. Springer International Publishing. https://doi.org/10.1007/978-3-319-62530-0_5

Derek McColl and Goldie Nejat. 2014. Recognizing emotional body language displayed by a human-like social robot. *International Journal of Social Robotics* 6, 2 (2014), 261–280.

Masahiko Mikawa, Masahiro Yoshikawa, Takeshi Tsujimura, and Kazuyo Tanaka. 2009. Librarian robot controlled by mathematical aim model. In *2009 ICCAS-SICE. IEEE*, 1200–1205.

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

Masahiro Mori, Karl F MacDorman, and Norri Kageki. 2012. The uncanny valley [from the field]. *IEEE Robotics & Automation Magazine* 19, 2 (2012), 98–100.

Michael G Morris and Viswanath Venkatesh. 2000. Age differences in technology adoption decisions: Implications for a changing work force. *Personnel psychology* 53, 2 (2000), 375–403.

Robin R Murphy. 2004. Trial by fire [rescue robots]. *IEEE Robotics & Automation Magazine* 11, 3 (2004), 50–61.

Daniel E O'Leary. 2019. GOOGLE'S Duplex: Pretending to be human. *Intelligent Systems in Accounting, Finance and Management* 26, 1 (2019), 46–53.

Guglielmo Papagni and Sabine Koeszegi. 2020. Interpretable Artificial Agents and Trust: Supporting a non-Expert Users Perspective. In *Culturally Sustainable Social Robotic*, M. Nørskov, J. Seibt, O. S. Quick (Eds.). (2020), IOS Press, Amsterdam. 653–662.

Guglielmo Papagni and Sabine Koeszegi. 2021a. A Pragmatic Approach to the Intentional Stance Semantic, Empirical and Ethical Considerations for the Design of Artificial Agents. *Minds and Machines* 31, 4 (2021), 505–534.

Guglielmo Papagni and Sabine Koeszegi. 2021b. Understandable and trustworthy explainable robots: A sensemaking perspective. *Paladyn, Journal of Behavioral Robotics* 12, 1 (2021), 13–30.

Charles Sanders Peirce. 1997. *Pragmatism as a principle and method of right thinking: The 1903 Harvard lectures on pragmatism.* SUNY Press.

Emily Pronin. 2009. The introspection illusion. *Advances in experimental social psychology* 41 (2009), 1–67.

Daniel B Quinn, Richard Pak, and Ewart J de Visser. 2017. Testing the efficacy of human-human trust repair strategies with machines. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 61. SAGE Publications Sage CA: Los Angeles, CA, 1794–1798.

Rafael Ramos-Garijo, Mario Prats, Pedro J Sanz, and Angel Pasqual Del Pobil. 2003. An autonomous assistant robot for book manipulation in a library. In *SMC'03 Conference Proceedings. 2003 IEEE International Conference on Systems, Man and Cybernetics. Conference Theme-System Security and Assurance (Cat. No. 03CH37483)*, Vol. 4. IEEE, 3912–3917. doi: 10.1109/ICSMC.2003.1244499

Paul Robinette, Ayanna M Howard, and Alan R Wagner. 2017. Effect of robot performance on human–robot trust in time-critical situations. *IEEE Transactions on Human-Machine Systems* 47, 4 (2017), 425–436.

Cynthia Rudin. 2018. Please Stop Explaining Black Box Models for High Stakes Decisions. *arXiv* (Nov 2018). arXiv:1811.10154v1 https://arxiv.org/abs/1811.10154

Selma Sabanovic, Marek P Michalowski, and Reid Simmons. 2006. Robots in the wild: Observing human-robot social interaction outside the lab. In *9th IEEE International Workshop on Advanced Motion Control*, 2006. IEEE, 596–601.

Fatai Sado, C Kiong Loo, Matthias Kerzel, and Stefan Wermter. 2020. Explainable goal-driven agents and robots-a comprehensive review and new framework. *arXiv preprint* arXiv:2004.09705 180 (2020).

Raymond Sheh. 2017a. Different XAI for different HRI. In *AAAI Fall Symposium-Technical Report.* 114–117.

Raymond Ka-Man Sheh. 2017b. "Why Did You Do That?" Explainable Intelligent Robots. In *Workshops at the Thirty-First AAAI Conference on Artificial Intelligence*, 628–634.

Keng Siau and Weiyu Wang. 2018. Building trust in artificial intelligence, machine learning, and robotics. *Cutter Business Technology Journal* 31, 2 (2018), 47–53.

Krishna Sood. 2018. The ultimate black box: The thorny issue of programming moral standards in machines [Industry View]. IEEE Technology and Society Magazine 37, 2 (2018), 27–29.

Robert P Spunt, Meghan L Meyer, and Matthew D Lieberman. 2015. The Default Mode of Human Brain Function Primes the Intentional Stance. *Journal of cognitive neuroscience* 27, 6 (2015), 1116–1124. 9

MS Sreejith, Steffy Joy, Abhishesh Pal, Beom-Sahng Ryuh, and VR Sanal Kumar. 2015. Conceptual design of a wi-fi and GPS based robotic library using an intelligent system. *International Journal of Computer and Information Engineering* 9, 12 (2015), 2504–2508.

Paul Thagard. 1989. Explanatory coherence. *Behavioral and brain sciences* 12, 3 (1989), 435–502.

Andreas Theodorou, Robert H. Wortham, and Joanna J. Bryson. 2016. Why is my robot behaving like that? Designing transparency for real time inspection of autonomous robots. Paper presented at *AISB Workshop on Principles of Robotics*, Sheffield, UK United Kingdom. (Apr 2016).

A Narasima Venkatesh. 2019. Reimagining the future of healthcare industry through Internet of medical things (IoMT), artificial intelligence (AI), machine learning (ML), big data, mobile apps and advanced sensors. *International Journal of Engineering and Advanced Technology (IJEAT), 9, 1 (2019)*. http://dx.doi.org/10.2139/ssrn.3522960, 3014–3019.

Douglas Walton. 2006. Examination dialogue: An argumentation framework for critically questioning an expert opinion. *Journal of Pragmatic*s 38, 5 (2006), 745–777.

Douglas Walton. 2011. A dialogue system specification for explanation. *Synthese* 182, 3 (2011), 349–374.

Tong Wang. 2019. Gaining free or low-cost interpretability with interpretable partial substitute. In *International Conference on Machine Learning*. PMLR, 6505–6514.

Karl E Weick, Kathleen M Sutcliffe, and David Obstfeld. 2005. Organizing and the process of sensemaking. *Organization science* 16, 4 (2005), 409–421.

Eva Wiese, Giorgio Metta, and Agnieszka Wykowska. 2017. Robots as intentional agents: using neuroscientific methods to make robots appear more social. *Frontiers in psychology* 8 (2017), 1663.

Daniel A Wilkenfeld and Tania Lombrozo. 2015. Inference to the best explanation (IBE) versus explaining for the best inference (EBI). *Science & Education* 24, 9-10 (2015), 1059–1077.

Jeffrey C Zemla, Steven Sloman, Christos Bechlivanidis, and David A Lagnado. 2017. Evaluating everyday explanations. *Psychonomic bulletin & review* 24, 5 (2017), 1488–1500.

Visual and Physical Plausibility of Object Poses for Robotic Scene Understanding

Dominik Bauer, Timothy Patten, Markus Vincze

Abstract

Humans use the relations between objects in a scene to determine how they may interact with, grasp and manipulate them. For robots, such an object-based scene understanding not only allows interaction with objects but also allows humans to interpret the robot's perception and actions. To gain a higher-level understanding of an observed scene, knowledge of the objects' poses is crucial. The poses, when combined with 3D models of the objects, allow for easy derivation of the interactions between objects, enabling reasoning about occlusion, collisions, support and, finally, manipulation by the robot. However, most related work does not consider scene-level object interactions but rather focuses on finding the pose of a single object in a given frame. Object interactions are considered only to augment training data or in post hoc verification steps. In contrast, we show that such scene-level information should be exploited during the estimation of the object poses themselves. Our main assumption is that all object hypotheses need to be plausible in terms of their visual observation and the physical scene in which they exist. In this chapter, we present our work on investigating the exploitation of this visual and physical plausibility for robust, accurate estimation and understandable explanation of object poses.

Keywords

robot vision, object pose estimation, object pose refinement, hypothesis verification, explainability

1 Introduction

The ability of a robot to explain its actions – or reasons why it might have failed – is an important building block for establishing and maintaining human trust [Lomas et al. 2012; de Graaf and Malle 2017; de Graaf et al. 2018]. For example, interactive explanations are an effective way to gain a deeper understanding of the reasoning provided [Dunne et al. 2005; Walton 2007; Arioua et al. 2017; Madumal et al. 2019]. But to provide such interactive explanations, the robot must attain a thorough understanding of the scene it inhabits. This may include the scene's objects, their location and their relationship to one another, for example expressed as their class, pose and spatial relations, respectively [Naseer et al. 2018]. Moreover, such an understanding enables the robot to perform tasks, such as grasping and manipulating objects, in the first place [Srinivasa et al. 2010; Chitta et al. 2012; Tremblay et al. 2018].

We hypothesize that, for the robot to provide an effective explanation of its understanding of a scene and its interactions with it, it must resolve to human-understandable reasoning approaches, such as how well the robot's understanding visually aligns with its camera images or how physically plausible an object's pose would be in a simulation of its estimated scene. We conjecture that both the visual and physical plausibility of the robot's scene understanding must be jointly considered and we examine their application to the object pose estimation task.

https://doi.org/10.34727/2022/isbn.978-3-85448-052-5_4

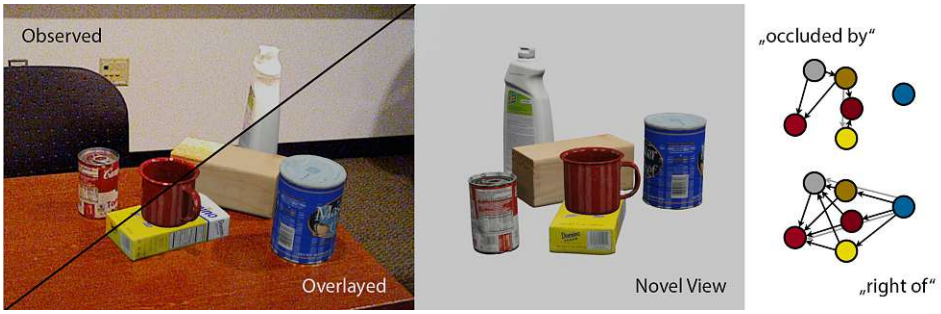


Figure 1 Applications of an object-based scene understanding. Left: Rendering the objects' models under their estimated poses allows to overlay and compare the robot's perception to the observed scene. Mid: Similarly, a novel view of the observed scene may be rendered. Right: Using the estimated poses, also the relations between the objects in the scene may be derived.

The poses, when combined with 3D models, allow the robot to manipulate the scene and explain it in terms of objects and their relations as illustrated in Figure 1.

This chapter provides an overview of our work exploring these hypotheses. In Section 2, we define visual plausibility through rendering and physical plausibility through simulation or evaluation of the static equilibrium. We present two different approaches for exploiting plausibility in object pose estimation. The methods we propose in Section 3 only require the 3D models of the objects and augment existing pose refiners. In Section 4, we propose novel object pose refinement methods based on reinforcement learning. These methods may jointly consider both aspects of plausibility that are discussed in this chapter. In Section 5, we present reasoning strategies that exploit this information for explanations in human-robot interaction. Finally, in Section 6, we discuss our findings and draw conclusions for future work.

2 Defining Visual and Physical Plausibility of Object Poses

A scene understanding represented by (semantically annotated) 3D models and their object poses allows to derive information about the scene that can be used for explanation and improvement of the poses themselves. For example, spatial relations between objects may be derived or a rendering of the estimated scene may be compared to the robot's camera image, as shown in Figure 1. Furthermore, the latter allows a robot to determine the plausibility of its scene understanding and subsequently explain why its actions might have succeeded or failed.

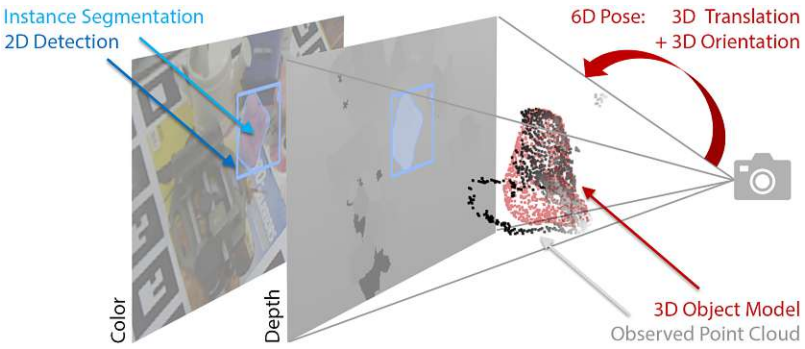


Figure 2 An object pose estimation pipeline. Left: A known object of interest is detected in the observed image. Mid: Using the instance segmentation mask, a cloud of all points predicted to belong to the object is generated from the corresponding depth image. Right: The task is to determine the 6D pose of the 3D model of the object such that it aligns to the observed image or point cloud.

The task of object pose estimation is to find the transformation T that aligns a 3D model of the object with its observation, as illustrated in Figure 2. We need to estimate this transformation by $\hat{T} = [\hat{R} \in SO(3), \hat{t} \in \mathbb{R}^3]$, i.e., a rotation $\hat{R}$ and a translation $\hat{t}$.

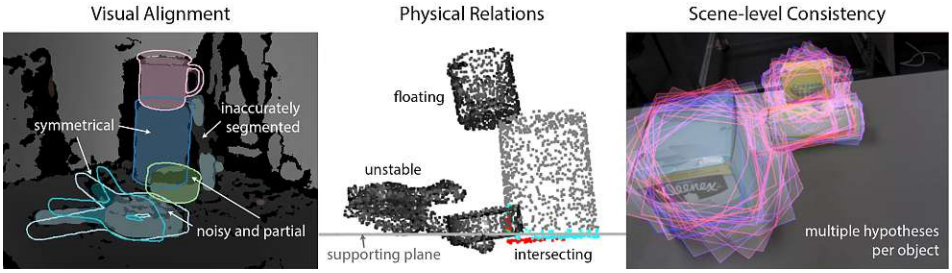


Figure 3 Challenges in object pose estimation. Left: Limited visibility, noise and inaccurate segmentation result in inaccurate pose estimates. Mid: The physical object relations in the estimated scenes violate the assumptions of plausible, static scenes. Right: Considering all scene-level interactions of multiple object under multiple (inaccurate) pose hypotheses quickly grows intractable.

The observation may be in the form of RGB or depth images. It is therefore only a partial and noise afflicted view of the object due to limited visibility from a single view and sensor limitations, as shown in Figure 3 (left). This problem is exacerbated in cluttered scenes and affects all parts of the perception pipeline – from detection, to segmentation and pose estimation. As a result, we might end up with multiple inaccurate pose hypotheses, as illustrated in Figure 3 (right). On the

one hand, to prevent failure, we want to verify and select the best available object pose before executing any robotic actions. On the other hand, we want to be able to explain why the robot selects a certain pose or why it decides that the pose is sufficiently accurate to base its interactions on it. In this section, we propose two approaches to this, based on visual alignment and physical plausibility.

2.1 Rendering-based Visual Plausibility

Object pose estimation and evaluation thereof are commonly based on the alignment of a 3D object model [Hodaň et al. 2020]. The Average Distance of Model Points (ADD) [Hinterstoisser et al. 2012] is the most used metric in related work. It measures the mean distance between corresponding model points $x \in X$ under estimated pose $\hat{T}$ and ground-truth pose T, or formally

$$ADD = \text{avg}_{x \in X} ||\hat{T}x - Tx||_2. \tag{1}$$

In contrast, the Visual Surface Discrepancy (VSD) [Hodaň et al. 2016, 2018], considers the discrepancy between the rendered depth images of the object under estimated pose $\hat{I}_d(\hat{T})$ and ground-truth pose $\hat{I}_d(T)$ by

$$VSD = \text{avg}_{p \in V(\hat{T}) \cup V(T)} \begin{cases} 0, & \text{if } p \in V(\hat{T}) \cap V(T) \text{ and } \Delta(p) < \tau, \\ 1, & \text{otherwise.} \end{cases} \tag{2}$$

The visibility under a given pose V is computed with respect to the observed depth image I_d and $\Delta(p)$ is the absolute difference between the rendered images at a pixel p.

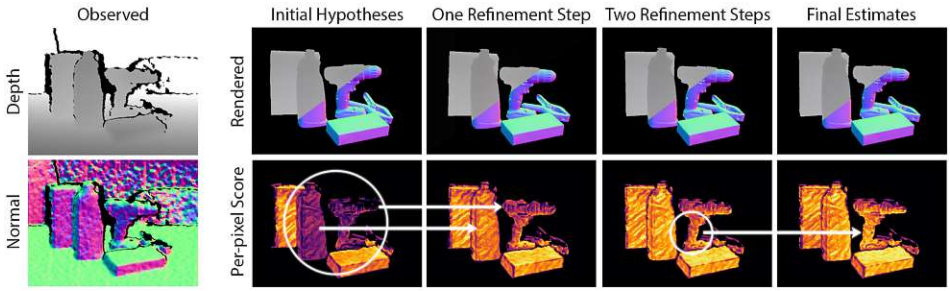


Figure 4 Example of the visual-alignment score. The observed depth and surface normals (left) are compared to the rendered objects under estimated pose (right). The resulting score for different sets of pose hypotheses (columns) is visualized below, where a more yellow color indicates better alignment with the observation. Adapted from [Bauer et al. 2022].

When estimating the pose of an object, the ground-truth pose is unknown and thus these metrics cannot be used to measure the quality of the pose estimate. Building on the idea of VSD, however, we suggest that the rendered view of a scene should be compared to the *observation* (i.e., the robot's camera view), as it can be considered a noisy version of the rendered object under the ground-truth pose T. If both align, we consider the estimate to be visually plausible. We define the visual-alignment score $\bar{a}$ in [Bauer et al. 2020c] that quantifies the average alignment between the object in the observed and rendered depth and normal images under the estimated pose $\hat{T}$. As illustrated in Figure 4, $\bar{a}$ is computed over all pixels with valid depth values, defined as $V = I_d > 0 \cup \hat{I}_d(\hat{T}) > 0$, by

$$\bar{a} = \frac{1}{2}\left(\text{avg}_{p \in V} a_d(p) + \text{avg}_{p \in V} a_n(p)\right), \tag{3}$$

with depth-based alignment a_d and normal-based alignment a_n per pixel p defined as

$$a_d(p) = \begin{cases} 1 - \frac{|d-\hat{d}|}{\tau}, & \text{if } |d-\hat{d}| < \tau \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

$$a_n(p) = \begin{cases} 1 - \frac{1-n\cdot\hat{n}}{\alpha}, & \text{if } 1 - n\cdot\hat{n} < \alpha \\ 0, & \text{otherwise,} \end{cases} \tag{5}$$

where $d \in I_d$ is the depth value and $n \in I_n$ is the corresponding normal at pixel p in the observation. The corresponding values in the rendered image are denoted by $\hat{d} \in \hat{I}_d(\hat{T})$ and $\hat{n} \in \hat{I}_n(\hat{T})$. The parameters τ and α limit the maximal admissible discrepancy.

2.2 Contact- and Simulation-based Physical Plausibility

Visual alignment alone may result in ambiguity under partial observability. We suggest that physical plausibility is able to resolve visually ambiguous cases. We define the physical plausibility of a scene as the combination of feasibility (non-intersecting, non-floating) and static stability of the objects therein, as illustrated in Figure 5.

Contact-based Formulation: We define these conditions based on two sets of critical points in [Bauer et al. 2020a], the intersecting points $\mathcal{I}$ and the contact points $\mathcal{C}$. These point sets depend on the signed distance δ between the object of interest and the scene. δ is computed for uniformly random sampled points $\hat{X}$ on the surface of the model X under an estimated pose $\hat{T}$. We compute these point sets with respect to a slack variable ε, accounting for inaccuracy due to the mesh representation and random sampling.

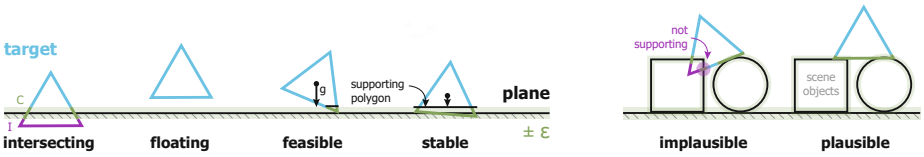


Figure 5 Definition of physical plausibility based on critical points for a single object (left) and a scene (right). If feasible, the center of mass projected in gravity direction must intersect the support polygon (convex hull of supported points) to be considered stable. Reprinted from [Bauer et al. 2022].

Intersecting points lie inside the scene objects' surface and contact points are within a small distance from them. Formally, we define

$$\mathcal{I} = \{\hat{x} \in \hat{X} : \delta(\hat{x}) < -\varepsilon\}, \tag{6}$$

$$\mathcal{C} = \{\hat{x} \in \hat{X} : |\delta(\hat{x})| < \varepsilon\}. \tag{7}$$

Based on these point sets we define an object to be

$$\text{not floating, if} \quad |\mathcal{C}| > 0, \tag{8}$$

$$\text{not intersecting, if} \quad |\mathcal{I}| = 0 \tag{9}$$

and feasible, if both conditions are satisfied.

Additionally, we consider the stability of the object, i.e., we determine whether it would be in static equilibrium (SE) under the estimated pose $\hat{T}$. To be in SE [Del Prete et al. 2016; Hauser et al. 2018], the object must satisfy the conditions of

$$\text{force balance} \quad \sum_i f_i + f_{ext} = \sum_i f_i + mg = 0, \tag{10}$$

$$\text{torque balance} \quad \sum_i (c_m - \hat{x}_i) \times f_i = 0 \quad \text{and} \tag{11}$$

$$\text{admissible contact force} \quad f_i \in \mathcal{K}, \tag{12}$$

where m is the mass of the object, c_m its center of mass, f_i is the contact force at contact point $\hat{x}_i \in \mathcal{C}$ and $\mathcal{K}$ is a friction cone.

The stability constraints may be approximated using the "support polygon principle" [Or and Rimon 2010]. The *support polygon* is defined as the convex hull of the projection of the contact points $\mathcal{C}$ onto the supporting plane. If the projection of the center of mass falls within the support polygon, the object is considered to be in SE [Or and Rimon 2010; McGhee and Frank 1968].

In static cluttered scenes (where gravity is the only external force acting upon objects), certain contact points of an object may not provide support in the gravity direction. Thus, they may result in an overestimation of its static stability, as the support polygon is enlarged by those contacts. Hence, as a compromise between the simplicity of the support polygon principle and the accuracy of solving for conditions (10)–(12), we consider the support polygon with respect to the *supported* points defined [Bauer et al. 2022] as

$$\mathcal{S} = \{\hat{x} \in C : \frac{n_{y(\hat{x})} \cdot g}{||n_{y(\hat{x})}||\,||g||} < 0\}, \tag{13}$$

where $y(\hat{x})$ is the closest point to $\hat{x}$ in the scene and $n_{y(\hat{x})}$ is its surface normal. Therefore, only the subset of contacts is considered onto which a force may be exerted in gravity direction g. See Section 4.2 for an application of this contact-based definition.

Simulation-based Formulation: Instead of evaluating physical plausibility based on contact points, we may also initialize the estimated scene in a physics simulation and evaluate its dynamic progression over time. Intuitively, a plausible configuration of a static scene should not be subject to any change due to gravity in the simulation. Since the 3D models used in the simulation and their physical parameters are inherently approximates of the real objects, we will observe at least slight displacement. Hence, rather than determining *whether* an object moved within the simulation, we want to determine *by how much* it moved over a (varying) period of time. To determine a stable pose, for example, we may want to simulate until the object no longer moves. In the simulation, resolving intersections typically generates an impulse that displaces the involved objects, causing the scene to “explode” in the worst case. To deal with estimated poses that result in intersecting objects, we might only simulate for a few steps at a time before setting the objects’ velocities back to 0 again. See Section 3 for an application of this simulation-based definition.

3 Enforcing Plausibility through Rendering and Simulation

To consider new objects, the methods presented in this section only require 3D models through using rendering and physics simulation . The proposed approaches enforce plausibility, exploit it to limit the search space given multiple pose hypotheses and improve initial poses. In Section 3.1, we present a simple approach for exploiting simulation for pose estimation. In Section 3.2, we present an integrated approach for improving refinement and augmenting it by verification.

3.1 Stable Object Pose Estimation

A simple proof-of-concept pose estimator [Bauer et al. 2020b] demonstrates the predictive power of considering plausibility for this task. It assumes only approximate object meshes and segmentation masks to be given; no additional training is required for pose estimation. This allows us to consider novel instances more easily than with end-to-end trained estimators. We derive a small set of physically plausible poses per object through physics simulation and clustering. Using the visual-alignment score, we are able to determine the visually most plausible candidate.

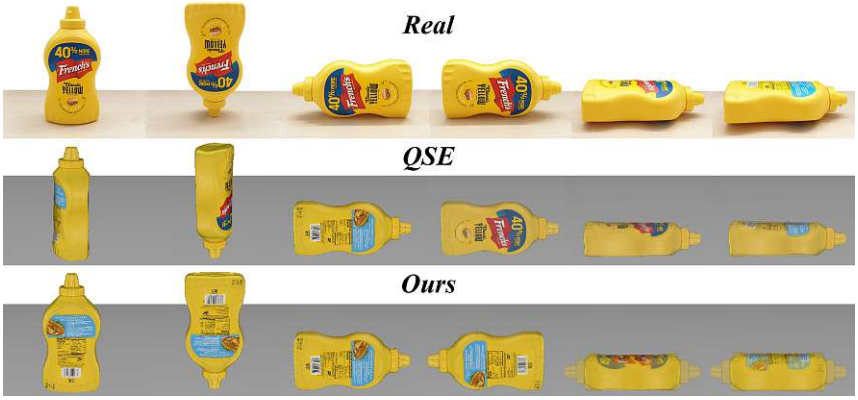


Figure 6 Stable object poses. Top to bottom: The real object, QSE [Goldberg et al. 1999] and our approach for isolated objects (ours). Multiple representatives of the same stable pose are transparently overlayed for QSE and ours. Reprinted from [Bauer et al. 2020a].

To determine the stable poses of an object, it is initialized under a uniformly random rotation in a physics simulator and dropped onto a plane. This assumption is motivated by the observation that objects in static scenes typically rest on horizontal planes, such as tables or shelves. Alternatively, more complex simulation scenes may be used for this purpose. Once the simulated object no longer moves, it has reached a stable pose. This process is repeated multiple times to sample a large number of potential stable poses. However, the resulting poses are highly redundant. First, multiple poses represent the same stable pose, albeit under in-plane rotation. Second, the object resting on different neighboring faces of the locally planar 3D model introduces a slight pose variance. To prune these superfluous poses, we discard in-plane rotation and cluster potential stable poses based on their angular distance. Each resulting stable pose represents the mean rotation and z-translation per cluster, with the plane normal defining the z-axis. Figure 6 shows a a comparison with the related probabilistic *quasi-stable estima-*

tion (QSE) approach [Goldberg et al. 1999] and real-world observations. While both our approach and QSE are able to reliably find all stable poses of an object resting on a horizontal plane, ours leverages a more general simulation-based approach. This would allow us to consider geometrically more complex simulation scenes or further physical properties of the object, beyond its shape and center of mass as in QSE.

To determine the pose of this object in an observation, we generate a pool of stable pose hypotheses by uniformly sampling in-plane rotations for each stable pose. Note that these hypotheses are inherently physically plausible for planar support. Given a segmented depth observation of the object, we may moreover estimate its in-plane translation as an offset from the rendered hypothesis. Among this pool of physically-plausible pose hypotheses, we need to find the visually most plausible pose. This is achieved by computing the visual-alignment score (3) for each hypothesis.

		simulation			
		$\tilde{C}_1$	$\tilde{C}_2$	$\tilde{C}_3$	$\tilde{C}_4$
visual	$\tilde{O}_1$	**51.5**	50.8	48.3	49.1
	$\tilde{O}_2$	51.6	**50.7**	48.4	49.0
	$\tilde{O}_3$	51.4	50.4	**47.8**	48.4
	$\tilde{O}_4$	48.9	48.6	45.2	**44.5**

Table 1 Influence of approximate object meshes on the *visual*-alignment score and *simulation*-based hypotheses generation. Results indicate the AR metric on Occluded LINEMOD.

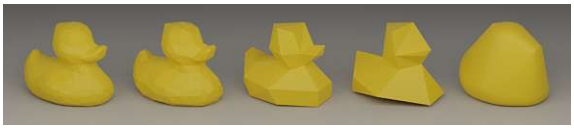


Figure 7 Approximate *duck* models $\tilde{O}_i$ with 704, 352, 70 and 34 faces and the convex hull $\tilde{C}$ of the full-resolution mesh $\tilde{O}$. Reprinted from [Bauer et al. 2020b].

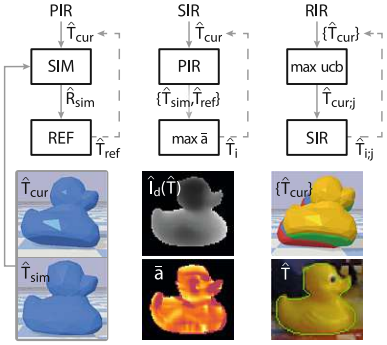
This simple approach achieves competitive pose accuracy on LINEMOD [Hinterstoisser et al. 2012] and Occluded LINEMOD [Brachmann et al. 2016], while also offering a general method to consider novel objects for pose estimation as it only depends on non-textured object meshes. To highlight the robustness of this approach, Table 1 shows our results on the Occluded LINEMOD dataset [Brachmann et al. 2016] using approximations of the object meshes as shown in Figure 7. We evaluate the impact of using the decimated meshes $\tilde{O}_i$ on the visual-alignment score (3) and the influence of using their convex hulls $\tilde{C}_i$ for the simulation-based stable pose generation. The decimated meshes are generated in Blender using the *decimate-collapse* operation. The reported Average Recall (AR) [Hodaň et al.

2020] is computed using the full-resolution object mesh and thus is solely dependent on the accuracy of the estimated pose. As shown per column in Table 1, our hypothesis scoring approach is highly robust to the decimated meshes, producing similarly accurate poses using the first three approximations. Shown per row, the stable pose hypotheses generated using our approach become increasingly inaccurate when the approximated resting shapes deviate farther from the original shape, i.e., with approximations $\tilde{C}_3$ and $\tilde{C}_4$.

3.2 Integrated Object Pose Refinement and Verification

An important step in object pose estimation pipelines is pose refinement. In pipelines yielding multiple pose hypotheses, the best hypothesis must be selected through pose scoring. Moreover, we want to verify the plausibility of the estimated object pose when using it for robotic manipulation, leveraging the pose scoring. With VeREFINE [Bauer et al. 2020c], we integrate iterative refinement, physics simulation and visual-alignment scoring in a joint optimization. We evaluate this approach on pose estimation datasets and in real-world grasping experiments.

(a) Initial pose estimates in the simulation environment (top) are improved using VeREFINE (mid, bottom), enabling successful robotic grasping.



(b) PIR: Integration of physics simulation (SIM) and iterative refinement (REF). SIR: Supervision using verification score $\bar{a}$. RIR: Regret minimization.

Figure 8 Grasping YCB objects with a Toyota HSR (a) and the iterative approaches proposed in VeREFINE (b), given an initial object pose estimate ($\hat{T}_{cur}$). Adapted from [Bauer et al. 2020c].

During refinement, we would like both discussed aspects of plausibility to inform one another. We achieve this by interleaving physics simulation steps with iterative refinement steps, as illustrated in Figure 8b (Physics-guided Iterative Refinement, *PIR*). Thereby, simulation guides refinement towards physically more

plausible poses, while alignment-based refinement improves visual plausibility. Both steps work complementary, improving each other's initialization.

However, either step might diverge, for example, due to bad initialization. The simulated object might topple over and move away from its true pose. Local refinement may determine incorrect correspondences and move toward a false pose. To contain these issues, we embed the visual-alignment score (3) in the refinement loop, as shown in Figure 8b (Supervised Iterative Refinement, *SIR*). Note that this also facilitates pose verification.

Generally, we might have to refine more than one object pose hypothesis. For example, with the pose estimator proposed in Section 3.1, multiple in-plane hypotheses need to be considered per stable pose hypothesis. With a growing number of hypotheses, simply refining and scoring all of them becomes computationally expensive. Rather, we want to spend a fixed budget of refinement iterations. We propose to consider the efficient allocation of the refinement budget as a multi-armed bandit problem. To minimize the regret of choosing to refine a sub-optimal hypothesis with respect to its visual-alignment score, we employ the Upper Confidence Bound policy (UCB) [Auer et al. 2002], as shown in Figure 8b (Regret-minimizing Iterative Refinement, *RIR*). The policy balances exploitation of high-scoring hypotheses with exploration of alternative, potentially better hypotheses.

We extend our approach to multiple objects per scene, considering the scene-level interactions of objects. We cluster scene objects based on their support relationships, with each cluster starting from a base object in contact with the supporting plane. The clusters are then ordered from front to back, i.e., starting from the least occluded base object. To yield physically plausible configurations, we iteratively add objects from the ordered clusters to the simulated scene during refinement. Each object's pose hypotheses are refined as before, albeit considering the visual plausibility of the whole scene. The highest scoring hypothesis per object is added to the simulation scene used for the subsequent objects, allowing the consideration of occlusions and support relationships between them.

Table 2 shows the results of the different single- and multi-hypotheses approaches we propose in VeREFINE [Bauer et al. 2020c] on the YCB-Video dataset [Xiang et al. 2018]. The dataset contains scenes of 3-6 YCB objects [Calli et al. 2015], that are occluded and stacked upon each other in clutter. Initial pose hypotheses are generated using DenseFusion (DF) [Wang et al. 2019] and its associated refinement network (DF-R) is used as implementation of *REF* (see Figure 8b). Figure 9 depicts an ablation study to show the influence of the initial rotation and translation error as wel as the impact of partial depth data. For these

	AR	#ref/obj
DF-R	73.9	2
PIR	74.7	2
SIR	76.5	2
VF$_b$	77.6	10
VF$_d$	**77.8**	10

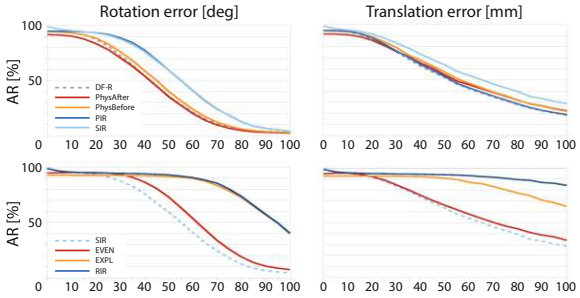
(a) Comparison on YCB-VIDEO.

	mustard	spam	foam	jello	banana	success	found	#ref/obj
DF-R	10	3	1	7	0	42%	46%	2
SIR	9	7	2	7	0	50%	70%	2
DF-R	10	6	5	9	1	62%	70%	10
MCTS	9	10	2	6	0	54%	78%	10
RIR	**10**	**10**	**9**	**10**	**4**	**86%**	**90%**	10

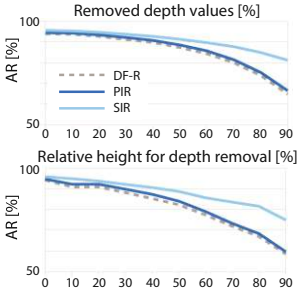
(b) Results of grasping experiments in percentage of *found* collision-free grasp poses and *success*ful grasp attempts.

Table 2 Evaluation of the methods in VeREFINE [Bauer et al. 2020c] (bold). Initial poses from DenseFusion [Wang et al. 2019], sampled to 1/5 hypotheses per object and refined with a budget of two refinement iterations per hypothesis and object for a total of 2/10 iterations.

experiments, the initial poses are generated by adding a uniformly random error of varying magnitude on top of the ground-truth poses.



(a) Using single hypotheses (top) and five hypotheses (bottom). EVEN and EXPL use our verification score to determine the best estimate and PIR for refinement. PhysBefore and PhysAfter apply simulation before and after refinement.



(b) Robustness to missing depth values using a single hypothesis with a fixed error magnitude of 5mm and 5deg.

Figure 9 Ablations on LM. Average Recall (AR) [Hodaň et al. 2020] values are reported at 5mm/deg steps (a) and every 10% (b), respectively, and are linearly interpolated in between. Adapted from [Bauer et al. 2020c].

The integration of physics simulation in the iterative refinement loop (PIR) improves the achieved accuracy by providing better initialization in each step. In Figure 9a (top left) we see how alternative ways of combining simulation with refinement may even reduce the performance. The use of the visual-alignment score (SIR) significantly improves accuracy, as indicated in Table 2a. It also improves

the robustness to partial depth data, as shown in Figure 9b. Our motivation for using a multi-armed bandit formulation for considering multiple hypotheses (RIR) is to balance exploration of the different hypotheses with exploitation of known high-scoring hypotheses. In the extreme case, the former would spend the budget of refinement iterations evenly among hypotheses (EVEN), while the latter would use it to refine a single hypothesis (EXPL). Figure 9a (bottom row) shows the benefit of using multiple hypotheses and our regret-minimizing approach. These findings also transfer to real-world grasping experiments with a Toyota HSR and using the GRASPA layouts [Bottarel et al. 2020] for reproducibility, illustrated in Figure 8a. As indicated by the results in Table 2b, both our single hypothesis (SIR) and multi-hypothesis approaches (RIR) significantly improve grasp success compared to the baseline refiner (DF-R) and a competing approach that uses a combination of physics simulation and refinement in a Monte Carlo tree search (MCTS) scheme [Mitash et al. 2018].

4 Enforcing Plausibility in Learning-based Approaches

The methods presented in Section 3 consider the visual and physical aspects of plausibility separately. For example, in Section 3.2, enforcing physical plausibility through simulation competes with enforcing visual plausibility through iterative refinement, illustrated by the experiments in Figure 9a (top). Instead, the influence of both plausibility aspects should be dynamically adapted depending on the scene configuration and refinement state. We want to leverage the contact-based constraints (8)–(12) directly for refinement. This motivates the design of a learning-based, plausible pose refinement approach.

4.1 Reinforced Point Cloud Registration

As the first step in this direction, we propose a novel approach to the related task of point cloud registration [Bauer et al. 2021]. We pose the iterative registration task as determining a policy that selects basic registration actions in each step, as illustrated in Figure 10. Inspired by [Shao et al. 2020], we use discrete steps per axis, separately for rotation and translation. These actions, for example, translate the source by a small offset in x direction. Our registration agent (ReAgent) is trained using imitation and reinforcement learning. Its formulation allows the incorporation of additional constraints – such as physical plausibility – by including them in the agent's reward.

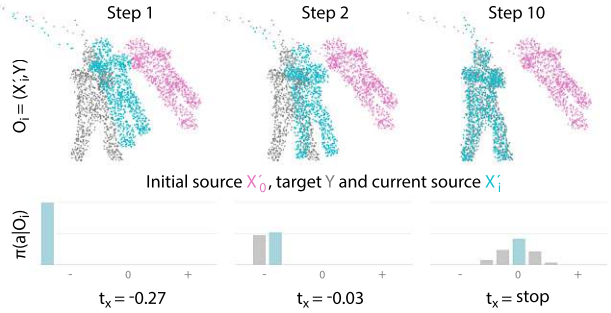


Figure 10 Iterative registration using ReAgent. The source point cloud (cyan) is aligned to the target point cloud (gray), starting from an initial source (magenta). ReAgent follows policy π by taking action $a_i = \arg\max_a \pi(a|O_i)$ given the current observation O_i, improving registration step-by-step. Reprinted from [Bauer et al. 2021].

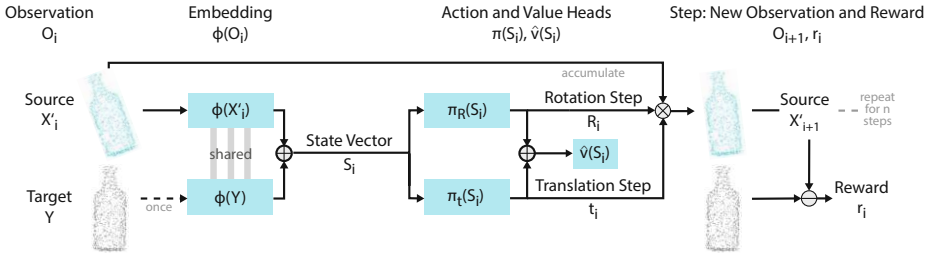


Figure 11 Architecture overview for one iteration of ReAgent. Reprinted from [Bauer et al. 2021].

The agent is implemented as a neural network, illustrated in Figure 11. The observed point clouds are embedded into a state space to reduce their dimensionality. The embedding uses a siamese PointNet-like architecture [Qi et al. 2017], generating a global feature that represents each point cloud. Two policy heads then predict the discrete distribution representing the policies for the rotation and translation action to be selected next. This process is also visualized in Figure 10 (bottom).

Since jointly learning the embedding and the registration policies from scratch using reinforcement learning (RL) might not converge (quickly), we opt for a hybrid training approach that also includes imitation learning (IL). Through IL, the agent should learn to replicate the behavior of an expert. We define an expert registration policy with perfect information (ground-truth transformation T) and, in each iteration, selects the actions that take the largest step toward alignment. Additionally, the agent is reinforced by a symmetry-aware point-cloud alignment

reward. The resulting loss is a combination of a cross-entropy loss for IL and the Proximal Policy Optimization (PPO) loss [Schulman et al. 2017] for RL.

	PoseCNN	DeepIM	Multi-ICP	ReAgent (IL)	ReAgent (IL+RL)
AD $< 0.10d$ (↑)	62.8	88.6	92.1	**98.7**	**98.7**
AD $< 0.05d$ (↑)	26.9	69.2	68.6	90.6	**91.1**
AD $< 0.02d$ (↑)	3.3	30.9	19.0	38.8	**39.6**

Table 3 Comparison of object pose refinement methods on LINEMOD (mean over per-class results) using PoseCNN [Xiang et al. 2018] for initial object pose and segmentation.

Figure 12 Qualitative examples on LINEMOD using ReAgent (IL+RL). In the top row, 1024 points are sampled within the estimated segmentation mask. The black box indicates the zoomed-in view. Outlines are shown for target (gray), initial (magenta) and current source pose (cyan). The last column shows a failure case. Reprinted from [Bauer et al. 2021].

In [Bauer et al. 2021], we show that our lightweight approach achieves faster inference as well as improved accuracy and robustness to noise and initialization as compared to related learning-based approaches on ModelNet40 [Wu et al. 2015] and ScanObjectNN [Uy et al. 2019]. Experiments on LINEMOD [Hinterstoisser et al. 2012], moreover, show high accuracy when applying ReAgent to the pose refinement task. Table 3 shows the comparison of our method to DeepIM [Li et al. 2018] and a rendering-based multi-hypothesis approach (Multi-ICP) [Xiang et al. 2018], employing initial poses and segmentation mask estimated using PoseCNN [Xiang et al. 2018]. When applied to the pose refinement task, our point cloud registration method achieves state-of-the-art performance on the LINEMOD dataset. The results obtained with tighter AD thresholds indicate the benefit of the combined IL and RL approach. Furthermore, Figure 12 illustrates the sampling of the source point cloud and qualitative examples of the accuracy of our ReAgent approach.

4.2 Reinforced Object Pose Refinement and Verification

When we apply the method from Section 4.1 (ReAgent) to cluttered scenes such as the ones observed in the YCB-Video dataset, we must cope with partial point clouds that may contain outliers from neighboring objects due to occlusion and inaccurate segmentation. Additionally, the initial pose estimates are affected by these challenges and are, in general, less accurate than in the single object case previously evaluated.

As we suggested in Section 2.2, additional consideration of physical plausibility allows us to resolve the resulting visual ambiguities. To this end, for *SporeAgent* [Bauer et al. 2022], we integrate our contact-based formulation from [Bauer et al. 2020a] with ReAgent. We modify it further to consider object symmetries, outlier points and visual-alignment scores. As a result, a learning-based approach similar to VeREFINE [Bauer et al. 2020c] (Section 3.2) is achieved that jointly considers both aspects of plausibility.

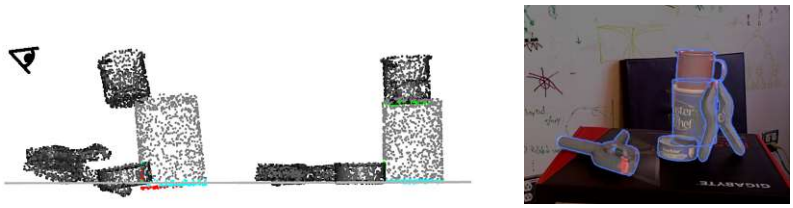


Figure 13 Initial scene representation (left) and refined poses using SporeAgent (mid and right). The critical points for one target object (gray) are shown – intersecting (red), contact (green) and supported (cyan). Adapted from [Bauer et al. 2022].

Physical plausibility is considered at two points in the refinement pipeline. First, we define an additional reward term that reinforces the agent to reach SE, approximated using the support polygon principle for the supported points $\mathcal{S}$ (as defined in Section 2.2). Second, we discover that the surface distance $\delta(\hat{x})$ is a useful input signal for the agent. It provides the underlying information required to determine the SE and, in addition, orients the object within the scene by including the distance to the supporting plane. As illustrated in Figure 13, these extensions allow the agent to resolve implausible configurations.

Visual plausibility with respect to the point clouds is already considered by the refinement itself. Additionally, to evaluate the iterative results, we leverage the visual-alignment score (3). Thereby, we are able to determine the overall most plausible (and accurate) object poses. This reduces the effect of the agent oscillating between two similarly fitting poses for fine alignment, as we observed in our

experiments, and allows to resort to the initial pose should the refinement diverge. Figure 4 shows a qualitative example for scoring.

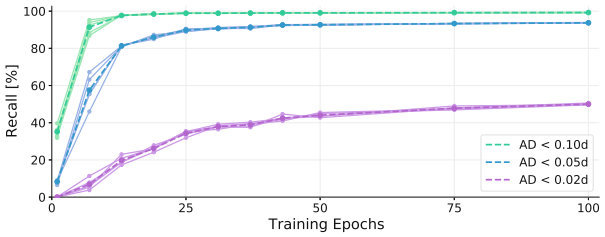
To further adapt the method to the task of object pose refinement in clutter, we introduce an outlier-removal subnetwork. Based on a concatenation of local and global features, this subnetwork is tasked with labeling geometrical outliers and is trained under an artificial segmentation error. The latter is an input augmentation that samples a coherent patch from the ground-truth segmentation mask, simulating occlusion and potentially including background pixels. The outlier predictions prune these geometrical outliers before the computation of the global feature used in the state vector (see Figure 11). Moreover, we adapt the expert policy to consider symmetrical objects by following the shortest trajectory to any symmetrical pose. To this end, we propose a canonical object frame in which the symmetry axes coincide with the origin, allowing symmetrical poses to be reduced to rotations. As a result, the symmetry-aware expert policy tends toward the symmetrical pose with minimal rotation from the current pose estimate.

	PoseCNN	ICC-ICP	P2Pl-ICP	w/ VeREFINE	Multi-ICP	SporeAgent
ADD AUC (↑)	51.5	67.5	68.2	70.1	77.4	**79.0**
AD AUC (↑)	61.3	77.0	79.2	81.0	86.6	**88.8**
ADI AUC (↑)	75.2	85.6	87.8	88.8	92.6	**93.6**

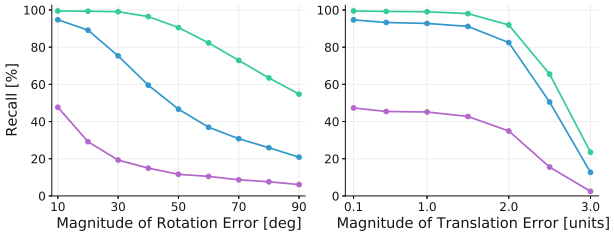
Table 4 Comparison of depth-based refinement methods on YCB-VIDEO (mean over per-class results) using PoseCNN [Xiang et al. 2018] for initial object pose and segmentation.

Table 4 shows the improved accuracy of SporeAgent compared to related depth-based refinement methods on YCB-Video [Xiang et al. 2018]. All compared methods use initial poses and segmentation masks estimated using PoseCNN [Xiang et al. 2018]. We compare our method to Iterative Collision Check with ICP (ICC-ICP) [Wada et al. 2020], vanilla Point-to-Plane ICP (P2Pl-ICP) [Chen and Medioni 1992; Zhou et al. 2018], P2Pl-ICP augmented by single-hypothesis VeREFINE [Bauer et al. 2020c] and a rendering-based multi-hypothesis approach (Multi-ICP) [Xiang et al. 2018]. While VeREFINE is able to significantly improve the results of the simple ICP approach by combining physics simulation with visual-alignment scoring, it is still inherently limited by the performance of the underlying refinement approach. In contrast, SporeAgent is able to exploit both sources of information to achieve state-of-the-art accuracy.

Figure 14a shows the training convergence of SporeAgent for five different random seeds on LINEMOD. For all evaluated thresholds, there is minimal variation in the recall beyond 50 epochs. Figure 14b shows an ablation study that highlights the robustness of SporeAgent to the quality of the initialization. For example, in



(a) Convergence of the mean recall per epoch (dashed) on LINEMOD for five random seeds (solid).



(b) Varying initialization error in rotation (left) and translation (right).

Figure 14 Ablations on LM. AD recalls with thresholds as fraction of the object diameter d [Hinterstoisser et al. 2012]. Reprinted from [Bauer et al. 2022].

the case of a translation error, the accuracy starts to decline only at a magnitude of around 2.0 units, which is limited by the number of iterations and the largest translation-step size.

5 Explaining Plausibility Violations

The consideration of plausibility offers not only a technical advantage but also supports users' understanding of the robot's perception and actions, thereby fostering trust. In [Papagni et al. 2021], we investigate how human interaction partners perceive plausibility-based explanations of robotic failure. Our proposed online study evaluates the impact of different explanation strategies on users' understanding of the robot and their trust in it after the interaction.

Participants in the study are instructed to assist a robot in locating and removing objects from a table, as shown in Figure 15 (top left). They are informed that their human-robot team may earn up to eight points in this task, one per object. This is to give the participants "something at stake" in the interaction. They are given a description of the next object to be be removed and are requested to provide the

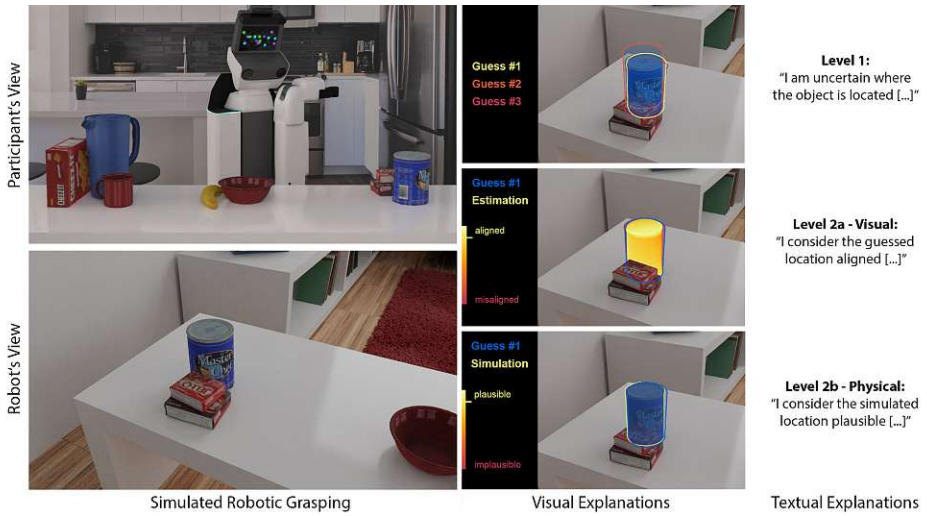


Figure 15 Rendered interaction from the view of the participants (top left) and the robot (bottom left). Example textual explanations are shown together with visualizations of uncertainty (top right), visual plausibility (mid right) and physical plausibility (bottom right). Adapted from [Papagni et al. 2021].

robot with an initial location. While hovering the cursor over the correct object, a circle indicates the corresponding location area. As a result, we aim to increase the perceived involvement of the participants in this human-robot interaction. After providing a location, they are shown rendered videos of the robot performing its task. Initially, robot (and hence *the team*) succeeds twice.

The third grasp attempt of the robot fails and the participants are shown different types of explanations, depending on the experimental condition to which they are assigned, as shown in Figure 15 (right). In a 2-by-2 study design, we modify the interactivity (*single-shot* or *multiple levels*) and the reasoning strategy (*visual alignment of the rendering* or *displacement in the physics simulation*) of the provided explanation. Participants then report their understanding of the explanation and answer short questionnaires regarding trust.

A technical pilot study has already highlighted the importance of the design of the visual explanations. Based on the findings of a currently ongoing user study, we will be able to further improve the visualizations and explanations provided by our object pose estimation approaches for deployment in human-robot interaction scenarios.

6 Conclusion and Future Work

This chapter discussed our definition of visual and physical plausibility, its technical benefit in object pose estimation and robotic grasping as well as its application in generating understandable explanations for human-robot interaction.

We showed that, by jointly considering these two aspects of plausibility, we are able to achieve increased pose accuracy in situations when each aspect alone would be ambiguous. We propose a set of object pose estimation and refinement approaches that are solely based on the 3D model of the objects and may be directly used to augment existing pipelines. Further exploiting the combined visual and physical plausibility information, we present a learning-based pose refinement method that considers the intersecting and supported points between interacting objects. Finally, we give an outlook on ongoing work investigating the exploitation of the plausibility information computed by our approaches to generate human-understandable explanations of robotic failure.

Nevertheless, many of the objects that robots have to deal with are not yet covered by the rigidity and static-scene assumptions of the proposed methods. Dealing with articulated (or even deformable) objects, potentially being manipulated by a human hand or robotic gripper and exposing high intra-class variance in texture and shape, is beyond the scope of this work. To this end, the visual plausibility considerations could be extended to include color information to deal with texture, thereby increasing the robustness of the methods to partial depth data. Considering, for example, hand-object contacts would allow the physical plausibility definition to be extended to these dynamic cases. Moreover, a robotic prototype that employs the presented methods to generate a scene explanation and the corresponding explanations of its actions (and failures) would allow for further evaluation of our approach in the ever-changing environments that the robots' human interaction partners inhabit.

Bibliography

Abdallah Arioua, Patrice Buche, and Madalina Croitoru. 2017. Explanatory dialogues with argumentative faculties over inconsistent knowledge bases. *Expert Systems with Applications* 80 (2017), 244–262.

Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. 2002. Finite-time analysis of the multiarmed bandit problem. *Machine Learning* 47, 2-3 (2002), 235–256.

Dominik Bauer, Timothy Patten, and Markus Vincze. 2020a. Physical Plausibility of 6D Pose Estimates in Scenes of Static Rigid Objects. *European Conference on Computer Vision Workshops*, 648–662.

Dominik Bauer, Timothy Patten, and Markus Vincze. 2020b. Scene Explanation through Verification of Stable Object Poses. *ICRA 2020 Workshop on Perception, Action, Learning.*

Dominik Bauer, Timothy Patten, and Markus Vincze. 2020c. VeREFINE: Integrating object pose verifi cation with physicsguided iterative refinement. *IEEE Robotics and Automation Letters* 5, 3, 4289–4296.

Dominik Bauer, Timothy Patten, and Markus Vincze. 2021. ReAgent: Point Cloud Registration using Imitation and Reinforcement Learning. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14586–14594.

Dominik Bauer, Timothy Patten, and Markus Vincze. 2022. SporeAgent: Reinforced Scenelevel Plausibility for Object Pose Refinement. *IEEE Winter Conference on Applications of Computer Vision*, 654–662.

Fabrizio Bottarel, Giulia Vezzani, Ugo Pattacini, and Lorenzo Natale. 2020. GRASPA 1.0: GRASPA is a robot arm grasping performance benchmark. *IEEE Robotics and Automation Letters* 5, 2 (2020), 836–843.

Eric Brachmann, Frank Michel, Alexander Krull, Michael Ying Yang, Stefan Gumhold, et al. 2016. Uncertainty-driven 6d pose estimation of objects and scenes from a single rgb image. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 3364–3372.

Berk Calli, Aaron Walsman, Arjun Singh, Siddhartha Srinivasa, Pieter Abbeel, and Aaron M Dollar. 2015. Benchmarking in manipulation research: Using the Yale-CMU-Berkeley object and model set. *IEEE Robotics and Automation Magazine* 22, 3 (2015), 36–52.

Yang Chen and Gérard Medioni. 1992. Object modelling by registration of multiple range images. *Image and Visual Computing* 10, 3 (1992), 145–155.

Sachin Chitta, E Gil Jones, Matei Ciocarlie, and Kaijen Hsiao. 2012. Mobile manipulation in unstructured environments: Perception, planning, and execution. *IEEE Robotics and Automation Magazine* 19, 2 (2012), 58–71.

Maartje MA de Graaf and Bertram F Malle. 2017. How people explain action (and autonomous intelligent systems should too). *AAAI Fall Symposium Series.*

Maartje MA de Graaf, Bertram F Malle, Anca Dragan, and Tom Ziemke. 2018. Explainable robotic systems. *Companion of the ACM/IEEE International Conference on Human-Robot Interaction*, 387–388.

Andrea Del Prete, Steve Tonneau, and Nicolas Mansard. 2016. Fast algorithms to test robust static equilibrium for legged robots. *International Conference on Robotics and Automation,* 1601–1607.

Paul E Dunne, Sylvie Doutre, and Trevor Bench-Capon. 2005. Discovering inconsistency through examination dialogues. *International Joint Conference on Artificial Intelligence,* 1680–1681.

Ken Goldberg, Brian V Mirtich, Yan Zhuang, John Craig, Brian R Carlisle, and John Canny. 1999. Part pose statistics: Estimators and experiments. *IEEE Transactions on Robotics and Automation* 15, 5 (1999), 849–857.

Kris Hauser, Shiquan Wang, and Mark R Cutkosky. 2018. Efficient equilibrium testing under adhesion and anisotropy using empirical contact force models. *IEEE Transactions on Robotics* 34, 5 (2018), 1157–1169.

Stefan Hinterstoisser, Vincent Lepetit, Slobodan Ilic, Stefan Holzer, Gary Bradski, Kurt Konolige, and Nassir Navab. 2012. Model based training, detection and pose estima-

tion of textureless 3d objects in heavily cluttered scenes. Asian Conference on Computer Vision, 548–562.

Tomáš Hodaň, Jiří Matas, and Štěpán Obdržálek. 2016. On evaluation of 6D object pose estimation. *European Conference on Computer Vision*, 606–619.

Tomáš Hodaň, Frank Michel, Eric Brachmann, Wadim Kehl, Anders GlentBuch, Dirk Kraft, Bertram Drost, Joel Vidal, Stephan Ihrke, Xenophon Zabulis, et al. 2018. BOP: Benchmark for 6D object pose estimation. *European Conference on Computer Vision,* 19–34.

Tomáš Hodaň, Martin Sundermeyer, Bertram Drost, Yann Labbé, Eric Brachmann, Frank Michel, Carsten Rother, and Jiří Matas. 2020. BOP Challenge 2020 on 6D Object Localization. *European Conference on Computer Vision Workshops* (2020).

Yi Li, Gu Wang, Xiangyang Ji, Yu Xiang, and Dieter Fox. 2018. Deepim: Deep iterative matching for 6d pose estimation. *European Conference on Computer Vision,* 683–698.

Meghann Lomas, Robert Chevalier, Ernest Vincent Cross, Robert Christopher Garrett, John Hoare, and Michael Kopack. 2012. Explaining robot actions. *ACM/IEEE International Conference on Human-Robot Interaction,* 187–188.

Prashan Madumal, Tim Miller, Liz Sonenberg, and Frank Vetere. 2019. A Grounded Interaction Protocol for Explainable Artificial Intelligence. *International Conference on Autonomous Agents and Multiagent Systems,* 1033–1041.

Robert B McGhee and Andrew A Frank. 1968. On the stability properties of quadruped creeping gaits. *Mathematical Biosciences* 3 (1968), 331–351.

Chaitanya Mitash, Abdeslam Boularias, and Kostas E Bekris. 2018. Improving 6D pose estimation of objects in clutter via physics-aware Monte Carlo tree search. *International Conference on Robotics and Automation,* 3331–3338.

Muzammal Naseer, Salman Khan, and Fatih Porikli. 2018. Indoor scene understanding in 2.5/3d for autonomous agents: A survey. *IEEE access* 7 (2018), 1859–1887.

Yizhar Or and Elon Rimon. 2010. Analytic characterization of a class of three-contact frictional equilibrium postures in three-dimensional gravitational environments. International Journal on Robotics Research 29, 1 (2010), 3–22.

Guglielmo Papagni, Dominik Bauer, Sabine Köszegi, and Markus Vincze. 2021. A Study Design for Evaluation of Trust and Understandability through Interactive Multi-Modal Explanations of Robotic Failure. *HRI 2021 WYSD Workshop.*

Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. 2017. Pointnet: Deep learning on point sets for 3d classification and segmentation. *IEEE/CVF Conference on Computer Vision and Pattern Recognition,* 652–660.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347* (2017).

Jianzhun Shao, Yuhang Jiang, Gu Wang, Zhigang Li, and Xiangyang Ji. 2020. PFRL: Pose-Free Reinforcement Learning for 6D Pose Estimation. *IEEE/CVF Conference on Computer Vision and Pattern Recognition,* 11454–11463.

Siddhartha S Srinivasa, Dave Ferguson, Casey J Helfrich, Dmitry Berenson, Alvaro Collet, Rosen Diankov, Garratt Gallagher, Geoffrey Hollinger, James Kuffner, and Michael Vande Weghe. 2010. HERB: A home exploring robotic butler. *Autonomous Robots* 28, 1 (2010), 5.

Jonathan Tremblay, Thang To, Balakumar Sundaralingam, Yu Xiang, Dieter Fox, and Stanley T Birchfield. 2018. Deep Object Pose Estimation for Semantic Robotic Grasping of Household Objects. *Conference on Robotic Learning,* 306–316.

Mikaela Angelina Uy, Quang-Hieu Pham, Binh-Son Hua, Thanh Nguyen, and Sai-Kit Yeung. 2019. Revisiting point cloud classification: A new benchmark dataset and classification model on real-world data. *International Conference on Computer Vision,* 1588–1597.

Kentaro Wada, Edgar Sucar, Stephen James, Daniel Lenton, and Andrew J Davison. 2020. Morefusion: multi-object reasoning for 6d pose estimation from volumetric fusion. IEEE/CVF *Conference on Computer Vision and Pattern Recognition,* 14540–14549.

Douglas Walton. 2007. Dialogical Models of Explanation. *Explanation-aware computing: Papers from the 2007 AAAI workshop.* Technical Report WS-07-06 (pp. 1–9). Menlo Park, CA: AAAI Press.

Chen Wang, Danfei Xu, Yuke Zhu, Roberto Martín-Martín, Cewu Lu, Li Fei-Fei, and Silvio Savarese. 2019. DenseFusion: 6D Object Pose Estimation by Iterative Dense Fusion. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 338–3347.

Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 2015. 3d shapenets: A deep representation for volumetric shapes. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1912–1920.

Yu Xiang, Tanner Schmidt, Venkatraman Narayanan, and Dieter Fox. 2018. Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. *Robotics: Science and Systems.*

Qian-Yi Zhou, Jaesik Park, and Vladlen Koltun. 2018. Open3D: A modern library for 3D data processing. *arXiv preprint arXiv:1801.09847* (2018).

Implementing Aspects of Trust in Robots

Design, Requirements, and Challenges of a Human-Robot Imitation System

Darja Stoeva, Margrit Gelautz

Abstract

Body motion is an important aspect in human-robot interactions. Giving robots the ability to imitate human motion can be beneficial for research on robot motion in a variety of applications. Furthermore, such a human-robot imitation system has the potential to provide a platform to investigate the facilitation of different types of trust in human-robot interactions. The goal of this paper is to describe the framework of a human-robot imitation system and investigate the system requirements imposed by different interaction settings. Several applications of imitation systems are discussed, along with their important characteristics and required features. Furthermore, open challenges for designing and developing human-robot imitation systems are discussed.

Keywords

human-robot interaction, body motion, imitation

1 Introduction

Because humans tend to assign social meaning to movements, human body movement is frequently perceived as expressive, making body motion an important component in social interactions. This tendency has been demonstrated in human interactions [Argyle 1975] as well as in situations where humans observed interactions between inanimate objects (moving geometrical shapes) [Heider and Simmel 1944]. Within the field of human-robot interaction, robot body movement and nonverbal behavior have been shown to influence how the robot is perceived in terms of animacy [Fukuda and Ueda 2010; Rosenthal-von der Pütten et al. 2018], anthropomorphism [Salem et al. 2013], feeling of co-presence [Krämer et al. 2016], and children's perceptions of a robot's warmth and competence [Peters et al. 2017].

Body movements designed for robots are often inspired by human body movements. It has been demonstrated that people prefer to interact with robots that exhibit human-like behavior over robots that exhibit machine-like behavior [Park et al. 2011]. Furthermore, when it comes to the dynamics of human-robot interactions, research has shown that people are more likely to coordinate their movement when interacting with a humanoid robot rather than a mechanical one [Chaminade et al. 2005] and when the motion is human-like rather than machine-like [Chaminade et al. 2008]. The dynamical feature of movement coordination is an important aspect of social interactions because it influences whether the interaction is perceived negatively or positively, which has a direct impact on the efficiency and stability of the interaction [Burgoon et al. 1995; Schmidt et al. 2012].

https://doi.org/10.34727/2022/isbn.978-3-85448-052-5_5

Human body movements are categorized with respect to their expressiveness by [Karg et al. 2013] as *communicative*, when they express or convey a message, *functional*, when they are used to accomplish a particular task, artistic, when they express a message in an exaggerated manner or when they are described as unfamiliar when compared to daily movements, or abstract, when they neither express a message nor serve a functional purpose. In the field of human-robot interaction, imitation, which is described as the ability of a robot to replicate human movement [Schaal 1999], serves as a promising tool to generate human-like movements. In principle, imitation systems provide a method to design and develop robotic body movements in any of the aforementioned categories of human movement. As a result, human-robot imitation systems can be used as an interactional framework to study body motion in human-robot interactions.

Furthermore, depending on the interaction setting for the system's intended application, human-robot imitation systems have the potential to provide a platform for studying different types of trust. There are two types of trust which would be applicable in this context, (1) interpersonal trust, which describes trust in social interactions based on the relationship that develops among the interactants [Ogawa et al. 2019], and (2) reliance trust, which is based on the belief that the robot will function as expected [Coeckelbergh 2012]. As a result, interpersonal trust can be studied in systems designed for social interactions, and reliance trust can be studied in systems designed for cooperative interactions. Using the system in various interaction settings may also allow for a comparison between these two different types of trust that can be facilitated in human-robot interactions.

The work presented here aims to propose an approach for the design and development of a human-robot imitation system with an intended application in mind. The main contribution is describing a framework for the design, development and evaluation of a human-robot imitation system. A second contribution is extending several existing applications of imitation systems, such as teleoperation and imitation learning, to also account for aspects such as interpersonal coordination, movement data collection, and exploration of body movements. In this context, we also include applications in the performing arts, which are not a very common point of interest in the field of robotics research. Finally, as a third contribution, a link between the envisioned applications and the system requirements of the proposed framework is established, which could aid the development process of future imitation systems. The paper is structured as follows. First, we describe the framework of a human-robot imitation system (Section 2), then we identify the application-dependent requirements of such a system for several potential applications (Section 3), followed by a discussion of open challenges (Section 4), and finally we provide a general conclusion (Section 5).

2 Framework of a Human-Robot Imitation System

Alissandrakis et al. [Alissandrakis et al. 2002] describe an agent-based perspective on the design of an imitation system that addresses five central questions: who, when, what, how to imitate, and how to evaluate the quality of imitation. The question of who to imitate refers to figuring out how to allow the robot to choose which interactant to imitate, especially in the case of multiple interactants. When to imitate refers to the times the robot needs to imitate and which movements within a behavior need to be imitated by the robot. Next, the system should consider what to imitate as part of an observed behavior, such as states, actions, and so on. How to imitate addresses the issue of mapping behavior from human embodiment to robotic embodiment. Finally, the question of how to evaluate the imitation is about finding a suitable metric to evaluate the similarity between the demonstrated and the resulting imitated behavior. Each of these questions has challenges and specific requirements depending on how the imitation system is intended to be used.

Such an agent-based approach is typically considered in the case of autonomous robots and aims to provide an approach independent of the robotic platform and the imitation task. In contrast, we argue in our work that the robotic platform, imitation task and system application all play an important role in the design and development process of a human-robot imitation system. Moreover, the majority of imitation systems considered in the literature are primarily aimed at applications of imitation learning or teleoperation. As opposed to that, in our research we extend the possible applications of an imitation system for humanoid robots and their usage scenarios while we lay out the framework of an imitation system from a developmental perspective.

The proposed framework for a human-robot imitation system is divided into three main components: (1) an *intended application* of the system, (2) a *technical implementation* with considerations based on the intended application, and (3) a suitable evaluation method based on the distinctive features of the imitation type. A flowchart of the suggested components for a human-robot imitation system is shown in Figure 1.

Because different interaction settings and tasks will have different system requirements, the intended application is one of the system's key components, making the technical implementation and method of evaluation application-dependent. Within the technical implementation, there are two important sub-components: (2.1) a means to sense human motion, and (2.2) a method that *translates the observed human motion into robot motion* (also shown in Figure 1).

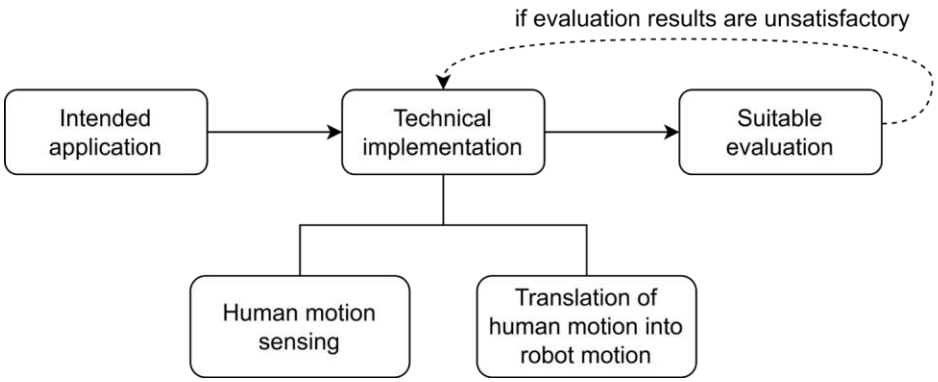


Figure 1 Flowchart showing the components of a human-robot imitation system

These two components are necessary in order to allow the robot to imitate human motion, and the choice of each of the components will affect the overall system performance. After the system has been implemented, a suitable evaluation should be carried out, which would be dependent on the specified system requirements and the distinctive features of the imitation type. Once the evaluation has been performed, depending on the results, certain improvements to the imitation system might be necessary, leading the development of such a system to undergo another cycle of (refined) technical implementation and evaluation. The subsections that follow go into greater detail about each of the components.

2.1. Intended Application

The first thing to consider when designing and developing a human-robot imitation system is the intended application. The importance of the application in the development of an imitation system has less to do with the end-goal and more to do with the interaction setting, which can be more interpersonal (e.g., mirroring) or more cooperative (e.g., teleoperation). The interaction setting is important because of the specific requirements that are required in various application contexts, which would define the necessary distinctive features of the imitation system. Some applications have stricter requirements while others provide more room for exploration. Additionally, the application also determines which methods are suitable candidates for system evaluation.

The potential applications of human-robot imitation systems vary depending on the interaction setting and the goal to be achieved with the imitation. Examples of such applications include *imitation learning* [Calinon and Billard 2007], *teleoperation* [Zuher and Romero 2012], *interpersonal coordination* (mirroring and synchrony) in social interactions [Hasumoto et al. 2020; Alibeigi et al. 2017],

movement data collection for interactive scenarios and expressive behavior (e.g., building datasets for nonverbal behavior), *exploration of body movements,* and the *performing arts* [Nakazawa et al. 2002]. Section 3 delves deeper into each of these applications in terms of technical requirements and important system evaluation features.

2.2. Technical Implementation

Following the selection of an application, the next step is to define the interaction settings, which will influence the method of human motion sensing in terms of joint positions, as well as the translation of human motion into robot motion in terms of converting joint positions to joint angles. The latter is required because the motor commands for robots are usually specified in terms of joint angles. In addition, to make the conversion from joint positions to joint angles feasible, the human joint positions need to be derived and processed in 3D space. The two most common methods for detecting 3D human joint positions are the use of motion capture systems such as Vicon[1] and the use computer vision algorithms for human pose estimation. Motion capture systems include camera based systems which comprise markers, attached to specific body parts (e.g., joints), and multiple cameras to track the markers and provide their positions in 3D space, or systems based on inertia sensors positioned on body parts without the need for external cameras. Computer vision algorithms are markerless pose estimators and provide joint positions directly in 3D space, such as the Kinect skeleton tracking module [Shotton et al. 2011], or estimate the 3D joint position from 2D body pose estimation [Mehta et al. 2017], or provide joint positions in the 2D camera space [Cao et al. 2019], which can then be used in combination with a depth-sensing camera to get the joints in 3D space (e.g., [Zabala et al. 2020]).

When choosing a method for sensing human motion it is important to consider the application scenario, which may impose different system requirements for obtaining the 3D human joint position data (depth-sensing camera, motion capture setups, or other means of sensing). The choice of method for motion sensing also includes the choice between usually more accurate sensing of 3D human joint positions in the case of motion capture systems, or allowing the human to move more freely in the case of using computer vision algorithms for human pose estimation.

Next, a model for converting joint positions to joint angles should be selected according to the system requirements imposed by the imitation goal of the application context. The imitation goal affects on the choice of the imitation type, which is closely connected to the method of system evaluation. The imitation

1 https://www.vicon.com/

type can aim to preserve the position of the end-effector with respect to the body [Zuher and Romero 2012], the pose for achieving body pose matching [Stoeva et al. 2021], or both [Alibeigi et al. 2017]. The choice of a model will depend on the imitation type but also the system's efficiency and allowed time for a delay in the imitated movement often need to be considered. The temporal aspect varies from aiming for real-time, to a specific tolerable time delay which can be further relaxed for offline applications.

Different approaches can be found for the model used to translate 3D human joint positions into robot joint angles. In the fields of robotics and mechanics, there are two main kinematic equations used for translating between 3D positions and angles: forward kinematics, which is the calculation of the 3D (end-effector) position given the joint angles, and inverse kinematics, which is the calculation of the joint angles given the 3D position. Methods for calculating the robot's joint angles based on the inverse kinematics include analytic and numeric solutions [Lynch and Park 2017; Craig 2005]. Numerical approaches are usually based on iterative algorithms that try to solve the inverse kinematics as an optimization problem (e.g., using the Jacobian). Analytical solutions, on the other hand, are usually approached in two ways, using geometry to find the angle between the links connecting two joints, or using algebra to express the angles in equations derived from forward kinematics. Both analytical and numerical approaches have advantages and disadvantages. For instance, numerical solutions are oftentimes much slower due to their inherent iterative nature and are highly dependent on the initial guess of joint angles. In contrast, even though analytical approaches provide closed-form solutions, it could be that they are too complex to manipulate into solvable equations. After choosing a suitable mode, the technical system can be considered complete and consists of two modules (2.1) a method for *sensing human motion* and (2.2) a model for *translating this motion to a robotic platform,* as depicted in Figure 1.

2.3. Suitable Evaluation

The next step in system design and development is to adequately choose a suitable method of evaluating the imitation system. Depending on the modules chosen during the implementation phase, it may be necessary to evaluate the accuracy of each module separately, before evaluating the full imitation system. For example, one approach is to evaluate the chosen model that translates the movements on how accurately it estimates joint angles from joint positions.

The first thing to consider for the full imitation system evaluation is the distinctive features of the imitation type, which are most commonly either the accuracy of the end-effector position with respect to the body, the body pose similarity

between the human pose and the imitated robotic pose, or both. The second factor to consider (if applicable) is the computational effort or time delay, which is the amount of time it takes for the imitation system to capture the human motion, translate it to robot motion, and send it to the robot as a motion command.

The evaluation methods can include quantitative, qualitative, a mix of both quantitative and qualitative, and subjective measurements. Quantitative measurements usually include the computation of the cosine similarity for the angular configuration of the pose [Guo et al. 2019; Zhang et al. 2018; Alibeigi et al. 2017], the mean squared error of the targeted versus actual joint positions [Guo et al. 2019; Zhang et al. 2018; Alibeigi et al. 2017], and the computation effort [Koenemann et al. 2014]. Qualitative measurements, on the other hand, usually include trajectory plotting of the X, Y and Z axis of the end-effector [Hirschmanner et al. 2019; Alibeigi et al. 2017; Mukherjee et al. 2015], plotting of the total error over time [Zhang et al. 2016; Koenemann et al. 2014], visual images, usually of motion sequences or specific postures, of the human posture and the robot exhibiting the imitated posture side by side [Guo et al. 2019; Zhang et al. 2018; Alibeigi et al. 2017; Zhang et al. 2016; Kim et al. 2016; Mukherjee et al. 2015; Ou et al. 2015]. Subjective measurements are typically based on user studies in which participants are asked to rate the quality of imitation by showing images or videos of the actual and imitated movement [Zuher and Romero 2012]. Depending on the intended application of the system and the imitation type, a suitable evaluation should be designed and performed. A good practice in evaluations of systems is to combine several methods of evaluation.

3 Application-dependent Requirements

As mentioned in the previous section, the development of a human-robot imitation system is highly dependent on the interaction settings of the system's application. Table 1 shows some of the potential applications for imitation systems with their system requirements, such as the methods of *human motion sensing,* the *imitation type,* the *time delay* between the performed and imitated movement, the evaluation features important for the evaluation process, and the *trust type* that can be facilitated and studied. In Table 1, the abbreviation "CV" stands for computer vision in the *human motion sensing column,* while in the *imitation type* column, "task-dependent" indicates that the choice of imitation depends on the targeted task, "both" stands for a compromise of preserving the end-effector position and body pose matching, and "any" stands for preserving any of the three imitation types explained in Subsection 2.2. The following subsections look into each of the suggested applications in relation to the aforementioned system requirements in the context of the interaction setting.

Applications	Human motion sensing	Imitation type	Time delay	Evaluation features	Trust type
Teleoperation	CV algorithms, motion capture	both	no delay	end-effector position, body pose similarity, time delay	reliance
Imitation learning	CV algorithms, motion capture	task-dependent	no delay	imitation feature, time delay	reliance
Interpersonal coordination	CV algorithms	body pose matching	from no to 5s delay	body pose similarity, time delay	interpersonal
Movement data collection	CV algorithms, motion capture	body pose matching	flexible	body pose similarity, time delay	reliance
Exploration of body movements	CV algorithms	any	no delay	open	interpersonal
Performing arts	CV algorithms, motion capture	any	flexible	open	reliance, interpersonal

Table 1 Potential applications of human-robot imitation systems with their system requirements and characteristics

3.1. Teleoperation

In situations in which the human operator cannot be physically present or in dangerous environments such as search and rescue, the method of robot teleoperation is envisioned as a possible approach [Penco et al. 2019; Koenemann et al. 2014; Stanton et al. 2012]. Due to the necessity of exact mapping of the human motion to the robot and the required high accuracy of the end-effector position with respect to the human body, an imitation system targeting teleoperation requires a high level of *human motion sensing accuracy*. For this application, a motion capture system usually provides more accurate readings than the use of available computer vision methods for the estimation of human pose. Motion capture often requires a specific interaction setting that typically includes several sensors or markers that need to be positioned on the human body, resulting in less spatial freedom and possibly discomfort for the interactant. This may not be an issue if the human and the robot are not interacting with each other face-to-face, which is usually the case for teleoperation. On the other hand, the accuracy of the involved human pose estimation algorithm has a significant impact on imitation performance when using computer vision methods. If computer vision is

the preferred method due to specific task requirements, the accuracy of the pose estimation algorithm can be evaluated using motion capture data as a reference. Since the idea behind robot teleoperation is for the human to be embodied in the robotic platform, the *imitation type* should preserve both end-effector position and body pose matching. The imitation should be performed with no *time delay* to allow for smooth control and quick feedback when controlling the robot. Thus, when evaluating an imitation system for teleoperation, the most important considerations for the *evaluation features* are end-effector position accuracy, body pose similarity metrics, and time delay. The *type of trust* that can typically be facilitated in this application is system reliance. For example, examining different types of teleoperation control and their influence on the trust of the system [Saeidi et al. 2017] or how different time delays affect the facilitated trust in the system [Rogers et al. 2017]. Ideally, for providing additional information to the human controller in order to ease the process of teleoperation, the imitation system should also include a virtual reality headset (e.g., [Hirschmanner et al. 2019]) and haptic force feedback (e.g., [Saeidi et al. 2017]).

3.2. Imitation Learning

The concept of using imitation learning (also known as learning from demonstration or programming by demonstration) as a method of teaching a robot to perform certain actions or behaviors stems from social learning in human interactions [Nehaniv and Dautenhahn 2007]. Researchers believe that robots capable of reproducing human movement could have advantages not only in allowing experts and non-experts to program behaviors for robots, but also as a means to better understanding of the concept of social learning [Breazeal and Scassellati 2002]. In an interaction setting where a robot needs to observe a human and learn specific behaviors, the important challenges to consider are how to successfully transfer the movement from the human to the robotic platform and which parts of the movement need to be reproduced. For such interaction settings, the use of motion capture or computer vision algorithms for human pose estimation is a common choice [Argall et al. 2009; Lee 2017]. However, to ensure that the required accuracy for imitation learning is met, both methods of *human motion sensing* should be evaluated in terms of achievable joint position accuracy. Because the interactant usually teaches the robot how to interact with the environment, in the context of imitation learning, the *imitation type* should usually preserve the position of the end-effector. However, in certain situations, depending on the task or behavior that needs to be imitated, it could be that both the end-effector position and body pose need to be maintained. Consequently, the choice of imitation type will depend on the task that needs to be completed or learned by the robot. In addition, the imitation should not have a noticeable

time delay between the interactant's demonstrated behavior and the imitated behavior by the robot. Similarly to teleoperation, the delay between the original and imitated motion is important for synchronized robot control. When controlling the robot to perform a particular task, immediate visual feedback is required when the motion needs to be corrected in appropriate time. This is especially important for novice users, and perhaps less so for experienced interactants as they may be able to adjust to how the system works more easily. The *evaluation features* that need to be considered for this application context should include methods for evaluating the accuracy of the imitation type and measurement of the time delay. As for the concept of *trust,* an imitation system for imitation learning can provide a platform for studying reliance trust, where the interactant would evaluate whether the system works as expected in both short and long term interactions. Another approach to studying trust in such systems is to investigate different methods of providing explanation about robot behavior and its effect on the facilitated trust in the system [Edmonds et al. 2019].

3.3. Interpersonal Coordination

When interacting socially with a robot, it is important for the interaction to be intuitive and smooth, meaning that both the human and the robot mutually influence and adapt to each other's behaviors. Interpersonal coordination, which includes mirroring and synchrony, is a phenomenon observed in human interactions as patterns that contribute to movement coordination and adaptation among interactants [Burgoon et al. 1995]. For *human motion sensing,* given the spatial restriction imposed by motion capture systems and the use of wearable markers or sensors, it might be preferable for interpersonal communication involving face-to-face interaction to rely on computer vision methods. This way, the interactant does not have to pay attention to the sensors/markers and will feel more comfortable to move and interact freely. Ideally, an internal (built-in) camera would be used, as no additional external equipment would be required. However, depending on where it is placed on the robot, the use of an internal camera has the potential to introduce further restrictions. Often cameras are positioned on a movable robot body part, for instance, the robot Pepper[2] has a depth camera placed in its head at the location of its 'eyes'. This can cause instability of the camera stream and, as a result, interfere with the data when the robot moves its head and perceives at the same time. When mirroring human motion, the system's *imitation type* should preserve body pose matching with the least amount of *time delay*. Compared to imitation learning and teleoperation, where the control of the robot requires no delay, for interpersonal coordination, the requirements on the mirror-

2 https://www.softbankrobotics.com/emea/en/pepper

ing behavior are more relaxed allowing for the time delay to range from no delay to 5 seconds. This time range comes from research in human interaction [Sato and Yoshikawa 2007; Louwerse et al. 2012], and it has also been investigated in human-robot interactions [Shimada et al. 2008]. Another significant difference from the applications of teleoperation and imitation learning is the complexity of interpersonal coordination within social interactions. In this case, the question of which body parts and when they should be imitated would need a greater consideration compared to imitation learning and teleoperation. The *evaluation features* for an interpersonal coordination system should use body pose similarity metrics and measurements of the time delay. Additionally, a user study can be designed to address the subjectivity of the perceived pose, which may include a collection of body pose similarity ratings as it was done in [V. Tuyen et al. 2018; Zuher and Romero 2012]. As interpersonal coordination usually manifests itself in social interactions, it provides a platform to study interpersonal *trust,* for instance how mirroring and synchrony behaviors affect the facilitated trust between the human and the robot. It is also important to note that privacy concerns arise in the context of social interactions. People who interact with the robot should be aware of any possible further usage of their data collected during the interaction.

3.4. Movement Data Collection

Translating human movement into robot movement is useful for designing and implementing body movements for interactive scenarios and expressive behavior for robots, especially nonverbal behavior. The ability to convert human motion into robot motion serves as a bridge and as a means for building datasets [Lee 2017] or potentially as a way to design expressive behavior for the targeted robotic platform [Fischer 2021; V. Tuyen et al. 2018; Liu et al. 2012; Häring et al. 2011]. The recorded and possibly annotated datasets can then be used to develop methods for recognizing and generating a nonverbal behavior of robots. Similar to teleoperation and imitation learning, the methods for *human motion sensing* can either rely on motion capture systems or computer vision algorithms for human pose estimation. In the best case scenario, for better recognition accuracy, the method of human motion sensing used to build the dataset should be the same as the one to be used in the application scenario. In order to generate human-like body movements, the *imitation type* in such systems should be body pose matching, so, as for interpersonal coordination, the *evaluation features* should include body pose similarity metrics and user studies. However, unlike the interaction setting for interpersonal coordination, in this case the interaction setting would not necessarily require a real-time interaction. Thus, there could be a more flexible requirement for the *time delay* between the human movement and the imitated movement by the robot. However, the *evaluation features* could also

include a measurement of the time delay. The observed delay could be a useful indicator of the overall system performance and allow for comparison with other imitation systems. The *type of trust,* in this case, would be system reliance, and a particularly interesting approach would be to study how the reliance on the system can have a feedback effect on the movement of the human being imitated.

3.5. Exploration of Body Movements

An imitation system could be useful for an overall exploration of the way the robot moves and getting a sense of its movement range, especially for novice users. Providing an interactive framework for movement exploration that relies on imitation could aid the interactant in understanding how the robot moves. This can support the creation of mental models of robotic behaviors and simulations of their movement capabilities. Furthermore, such a system could, under the supervision of a physical therapist, potentially be used in movement therapy, which usually consists of movement exercises (e.g., improvisation) designed to explore the physical capabilities of the human body [Halprin 2003]. Additionally, such a system can also be used as a way to promote social skills for individuals with autism spectrum disorder as it has been done in [Vallée et al. 2020; Boucenna et al. 2014]. For the application of body movement exploration, the person being imitated should be free to move around in space and interact in an unrestricted manner. Thus, similarly to interpersonal coordination, for *human motion sensing* the use of computer vision algorithms is preferable to motion capture setups. Because of the interaction setting, it is important that the imitation happens in real-time so that the observing-acting cycle is maintained. Accordingly, there should be no *time delay* in the movement imitation. Given the importance of how the body moves in this application, any of the three *imitation types* may apply, thus the *evaluation features* should be chosen accordingly. For body pose matching, the important feature for the evaluation would be the body pose similarity metrics, for preserving the position of the end-effector it would be the accuracy of the end-effector position. If the system is to be used in a therapeutic setting, it is also important to include experts (therapists) in the design and development process of the imitation system. For the *type of trust,* as the robot will play the role of an interactional partner with which an interpersonal trust can be facilitated, a possible investigation could be the link between trust and the success of movement therapy or improvement in social skills. Another approach could be to look into a possible relationship between the length of time spent interacting or moving with the robot and the facilitated interpersonal trust over time.

3.6. Performing Arts

A human-robot imitation system seems like an interactive platform that is likely to be an attractive tool for the performing arts. The reason for this is due to its ability to facilitate the processes of choreography development and performance preparation, among other things [Christiansen and Lindelof 2020]. Unlike teleoperation, imitation learning, and interpersonal coordination, which all have rather specific interaction setting and requirements, in the case of performing arts the approach is less restricted and allows for many different requirements to be considered. For instance, when the interaction setting is exploratory, the application of performing arts may have flexible requirements, but it can also have very strict requirements, as in choreographed dance. Therefore, the *imitation type* in a system with an envisioned application in performance could be approached in an experimental way, and the *evaluation features* would depend on the requirements of the artists interacting with the system, as well as the performance itself. The choice of a *human motion sensing* method would depend on the requirements and vision of the artist interacting with the system. The possibilities include a motion capture system or computer vision algorithm. However, it should be considered that performers often move their bodies in unpredictable and unconventional ways, for instance suddenly falling on the ground with full force. Thus, in those situations to avoid damaging wearable sensors or markers computer vision methods might be more favorable. In this case, the *time delay* between the performed and imitated movement is rather flexible, especially if the interaction setting is exploratory. When the imitation includes some delay, the artist may discovers interesting movement responses by the robot. Similar to the exploration of body movements with an imitation system, in the case of performing arts, the *imitation type* can be preserving body pose matching, the end-effector position, or both. The imitation system should be evaluated using *evaluation features* chosen according to the imitation type and the artist's requirements. When developing an imitation system for a specific artist, or performance preparation, it is important to include the artist or art director in the design and development process of the system. Regarding trust, there is potential for both reliance and interpersonal trust to be facilitated in the case of performance, again depending on the interaction setting.

4 Discussion

Body motion is an important ability that allows for the fulfillment of different types of actions. Enabling robots to use body motion as a way to communicate and interact with humans is a promising behavior for a fluid and intuitive HRI. With the high relevance and increased research interest in nonverbal behavior for the

design of future robots, it is important to consider which human behaviors are appropriate to adopt to robot behaviors. To allow for such research possibilities, we propose a framework of a human-robot imitation system in the simplest form that can serve as a foundation on which more complex behaviors can be developed. This is partly inspired by the works of [Jordanous 2020] and [Brooks 1991], which argue for incremental development of robot behavior, where each behavioral layer adds more complexity to the robot's capabilities.

Human-robot imitation systems have a wide range of applications from which many different research paths emerge. The future technical development of imitation systems highly depends on the advances in motion capture systems and robotic body design. From a broader perspective, building upon an imitation system has the potential to provide platforms for a better understanding of how artificial agents (robots) and humans exchange movements, how they differ from human interactions and how they can contribute to our understanding of body motion. In this spirit, human-robot imitation systems could also provide further insights into the role body motion has in human interactions.

Even though imitation behavior is a promising skill for social robots, current and potential future challenges must be considered. For the design and development of a human-robot imitation system we have identified several open challenges. In the following, we look in more detail into these challenges, which include the accuracy of sensing human motion, the correspondence problem of mapping the behavior from one body to another morphologically different body, the characteristics of the imitated motion, the choice of suitable evaluation metrics, and some ethical considerations when imitation systems are used in social interaction settings.

- **Accuracy of Human Motion Sensing**
  One of the challenges that arise when dealing with the requirement of sufficiently accurate imitation of human motion is choosing the appropriate method for human motion sensing. Due to the different characteristics of the currently available methods, there will be a trade-off between the availability, comfort handling and cost-efficiency of non-contact sensors and markerless methods (typically based on computer vision methods for human pose estimation) on the one hand, and high accuracy requirements (which are more easily met by motion capture devices) on the other. This compromise requires careful consideration of what is possible and what is necessary (in the case of special conditions) to meet the envisioned imitation goal. In addition, computer vision methods can introduce further challenges such as dealing with ambiguities, for example if there is more than one person in the camera view.
- **Correspondence Problem**
  Dealing with the physical differences and constraints of robots is another chal-

lenge in designing and developing a human-robot imitation system. The task of properly mapping the human motion to the robotic platform has been defined as the *correspondence problem* [Nehaniv and Dautenhahn 1998]. A common approach to facilitate the mapping between dissimilar bodies is the use of humanoid robots due to their morphology being similar to that of humans (head, arms, etc.). However, this only partially solves the problem, because humanoid robot joint usually have different degrees-of-freedom than human joints [Yamane and Murai 2016]. The physical morphology differences between humans and robots also creates challenges in different interaction settings. One possible solution would be to allow the interactant to change the type of imitation within the interaction whenever the interactant finds it necessary. This, of course, could change as the interactant gains more experience with the robotic platform, but it would be a useful approach for novice users to try out different imitation types and explore the capabilities of the robot. In addition, this would allow for a better understanding of the robot's imitation capabilities for the human embodying the robot, as well as a reduction in the difficulty of properly mapping the human to robot behaviors for the targeted goal of imitation.

- **Imitated Motion Characteristics**
  The following two challenges are identified when it comes to the characteristics of the motion reproduced by the robot, such as motion speed and smoothness. Many humanoid robots often move at a slower speed than humans, usually because of safety measures. Thus, if the human demonstrator moves faster than the robot's maximum speed there would be two possible options for approaching the speed of the imitated motion. The robot would either aim at imitating all poses within the motion sequence resulting in a delayed imitated movement, or skip some poses of the motion sequence to minimize the delay to near real-time imitation. Skipping some poses causes gaps in the imitated movement, implying that the robot will not reach all of the positions within the motion sequence as performed by the human. Second, smooth motion reproduction by minimizing motion jerkiness is still a feature that is being researched. To meet this challenge several methods have been proposed, such as pre-processing the data of the human motion [Luo et al. 2013], or post-processing the converted data to robot motion [Zhu et al. 2017]. Both pre-processing and post-processing the motion data usually includes filters (e.g. Kalman filter) that remove sensor noise and smooth the motion trajectory. However, finding a suitable method to smooth the motion trajectory remains an ongoing research topic.
- **Suitable Evaluation Metrics**
  Another open challenge that goes hand in hand with the correspondence problem is how to suitably evaluate an imitation system in terms of the success of the imitation. So far, there are many inconsistencies in the literature regarding the methods used to evaluate human-robot imitation systems, making it difficult

to compare systems to each other. One solution would be to provide a comprehensive set of evaluation metrics that can be applied selectively based on the distinctive features of the system, which would include a combination of quantitative, qualitative, and possibly subjective observational evaluation methods. In this context, it would also be important to define the imitation type and identify the aim of the imitation. The imitation goal can be to focus on the motion itself (e.g., how human-like the motion is) or the accomplishment of a specific task (e.g., the success of grasping an object). This will also determine which distinctive features will be the focus of the evaluation process. The goal is to find a suitable method that measures how successful the imitation is based on the goal of the imitation and the system's key features (e.g. imitation type, time component, etc.).

- **Ethical Considerations in Social Interaction Settings**
  When dealing with tracking of human data, it is important that privacy issues are taken into account and that people interacting with the technology are provided with transparent information on how their data is being used. Concerns have also been raised that imitation systems designed for social interactions, such as in the case of interpersonal coordination, may deceive interactants. This deception is described as deceiving interactants into thinking that the robot has more cognitive abilities than it does [Sharkey and Sharkey 2020]. However, the authors argue that not all deceptions are wrong as long as the deception does not cause any negative impact on the person or society in general. This distinction between wrong and not wrong deceptions is a topic of ongoing discussion in the fields of ethics and philosophy. On the other hand, findings in social psychology indicate that interpersonal coordination increases likability and rapport between interactants [Burgoon et al. 1995]. These findings may have an impact on how the way interpersonal coordination is transferred to be used in human-robot interactions. The ability of the robot to exhibit interpersonal coordination could be used to some advantage for the application or the stakeholders selling the robot, which could have a negative impact on the interactant. Thus, an open question from an ethical point of view is: How can we ensure that the imitation system is not used for the wrong deception of the interactants? And is it ethical (because of the possible deception) to allow robots to take part in social interactions and express interpersonal coordination with the interactants?

5 Conclusion

Imitation systems have a wide range of potential applications within the field of human-robot interaction. This paper proposes a method for designing and developing a human-robot imitation system in light of various application scenarios.

The following elementary system components are identified: intended application, technical implementation, and suitable evaluation. Each of these elements, as well as their interrelationships are described and discussed. Based on an examination of several potential applications, the interaction setting with its specific requirements is identified to be a key aspect to consider in system design. The interaction setting can range from having a higher interpersonal component (e.g., imitation for the purpose of interpersonal coordination) to having a higher cooperative component (e.g., imitation targeted for teleoperation) interaction settings. The system requirements that may emerge from the interaction setting have an important influence on decisions for the technical implementation, but also for choosing a suitable evaluation method. The interaction setting is also closely related to the possibility of facilitating different types of trust between the human and the robot. Finally, open challenges in developing human-robot imitation systems are discussed along with possible approaches as a way to tackle them. Further research should aim to better understand in what ways body motion contributes to the overall interaction between a human and a robot, and how it can be tested not only as a stand-alone capability but also in combination with other robotic social capabilities.

Bibliography

Mina Alibeigi, Sadegh Rabiee, and Majid N. Ahmadabadi. 2017. Inverse kinematics based human mimicking system using skeletal tracking technology. *Journal of Intelligent and Robotic Systems* 85 (2017), 27–45.

Aris Alissandrakis, Chrystopher L. Nehaniv, and Kerstin Dautenhahn. 2002. Imitation with ALICE: learning to imitate corresponding actions across dissimilar embodiments. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans* 32, 4 (2002), 482–496.

Brenna D. Argall, Sonia Chernova, Manuela Veloso, and Brett Browning. 2009. A survey of robot learning from demonstration. *Robotics and Autonomous Systems* 57, 5 (2009), 469–483.

Michael Argyle. 1975. *Bodily Communication.* International Universities Press.

Sofiane Boucenna, Salvatore Anzalone, Elodie Tilmont, David Cohen, and Mohamed Chetouani. 2014. Learning of social signatures through imitation game between a robot and a human partner. *IEEE Transactions on Autonomous Mental Development* 6, 3 (2014), 213–225.

Cynthia Breazeal and Brian Scassellati. 2002. Robots that imitate humans. *Trends in Cognitive Sciences* 6, 11 (2002), 481–487.

Rodney A. Brooks. 1991. Intelligence without representation. *Artificial Intelligence* 47, 1 (1991), 139–159.

Judee K. Burgoon, Lesa A. Stern, and Leesa Dillman. 1995. *Interpersonal Adaptation: Dyadic Interaction Patterns.* Cambridge University Press.

Sylvain Calinon and Aude Billard. 2007. Active Teaching in Robot Programming by Demonstration. In *Proceedings of the International Symposium on Robot and Human Interactive Communication (RO-MAN)*. 702–707.

Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2019. OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43, 1 (2019), 172–186.

Thierry Chaminade, David W. Franklin, Erhan Oztop, and Gordon Cheng. 2005. Motor interference between Humans and Humanoid Robots: Effect of Biological and Artificial Motion. In *Proceedings of the International Conference on Development and Learning.* 96–101.

Thierry Chaminade, Erhan Oztop, Gordon Cheng, and Mitsuo Kawato. 2008. From self-observation to imitation: Visuomotor association on a robotic hand. *Brain Research Bulletin* 75, 6 (2008), 775–784.

Henning Christiansen and Anja Lindelof. 2020. Robots on stage. *EAI Endorsed Transactions on Creative Technologies* 7, 25 (2020), 1–13.

Mark Coeckelbergh. 2012. Can we trust robots? *Ethics and Information Technology* 14 (2012), 53–60.

John J. Craig. 2005. *Introduction to Robotics: Mechanics and Control* (3rd ed.). Pearson Education.

Mark Edmonds, Feng Gao, Hangxin Liu, Xu Xie, Siyuan Qi, Brandon Rothrock, Yixin Zhu, Ying Nian Wu, Hongjing Lu, and Song-Chun Zhu. 2019. A tale of two explanations: Enhancing human trust by explaining robot behavior. *Science Robotics* 4, 37 (2019), eaay4663. doi.10.1126/scirobotics.aay4663

Sarah Fischer. 2021. *Design and Evaluation of Non-Verbal Cues for the Robot Pepper.* Master's thesis. Technical University of Vienna (TU Wien), Vienna, AT.

Haruaki Fukuda and Kazuhiro Ueda. 2010. Interaction with a moving object affects one's perception of its animacy. *International Journal of Social Robotics* 2 (2010), 187–193.

Wei Guo, Jianxin Chen, Ming Zhang, and Zhaolai Pan. 2019. Geometry Based LM of Robot to Imitate Human Motion with Kinect. In *Proceedings of the International Conference on Image, Vision and Computing (ICIVC)*. 695–700.

Daria Halprin. 2003. *The Expressive Body in Life, Art, and Therapy: Working with Movement, Metaphor and Meaning.* Jessica Kingsley Publishers.

Markus Häring, Nikolaus Bee, and Elisabeth André. 2011. Creation and evaluation of emotion expression with body movement, sound and eye color for humanoid robots. In *Proceedings of the International Conference on Robot & Human Interactive Communication (RO-MAN).* 204–209. 9

Ryosuke Hasumoto, Kazuhiro Nakadai, and Michita Imai. 2020. Reactive Chameleon: A Method to Mimic Conversation Partner's Body Sway for a Robot. *International Journal of Social Robotics* 12 (2020), 239–258.

Fritz Heider and Marianne Simmel. 1944. An Experimental Study of Apparent Behavior. *The American Journal of Psychology* 57, 2 (1944), 243–259.

Matthias Hirschmanner, Christiana Tsiourti, Timothy Patten, and Markus Vincze. 2019. Virtual Reality Teleoperation of a Humanoid Robot Using Markerless Human Upper Body Pose Imitation. In *Proceedings of the International Conference on Humanoid Robots (Humanoids)*. 259–265.

Anna Jordanous. 2020. Intelligence without Representation: A Historical Perspective. *Systems* 8, 3 (2020), 1–18.

Michelle Karg, Ali-Akbar Samadani, Rob Gorbet, Kolja Kühnlenz, Jesse Hoey, and Dana Kulić. 2013. Body Movements for Affective Expression: A Survey of Automatic Recognition and Generation. *IEEE Transactions on Affective Computing* 4, 4 (2013), 341–359.

Mingon Kim, Sanghyun Kim, and Jaeheung Park. 2016. Human motion imitation for humanoid by recurrent neural network. In *Proceedings of the International Conference on Ubiquitous Robots and Ambient Intelligence (URAI).* 519–520.

Jonas Koenemann, Felix Burget, and Maren Bennewitz. 2014. Real-time imitation of human wholebody motions by humanoids. In *Proceedings of the International Conference on Robotics and Automation (ICRA).* 2806–2812.

Nicole C. Krämer, Carina Edinger, and Astrid M. Rosenthal-von der Pütten. 2016. The effects of a robot's nonverbal behavior on users' mimicry and evaluation. In *Proceedings of the International Conference on Intelligent Virtual Agents.* 442–446.

Jangwon Lee. 2017. A survey of robot learning from demonstrations for human-robot collaboration. *arXiv preprint arXiv:1710.08789* (2017).

Chaoran Liu, Carlos T. Ishi, Hiroshi Ishiguro, and Norihiro Hagita. 2012. Generation of nodding, head tilting and eye gazing for human-robot dialogue interaction. In *Proceedings of the International Conference on Human-Robot Interaction (HRI)*. 285–292.

Max M Louwerse, Rick Dale, Ellen G Bard, and Patrick Jeuniaux. 2012. Behavior matching in multimodal communication is synchronized. *Cognitive Science* 36, 8 (2012), 1404–1426.

Ren C. Luo, Bo-Han Shih, and Tsung-Wei Lin. 2013. Real time human motion imitation of anthropomorphic dual arm robot based on Cartesian impedance control. In *Proceedings of the International Symposium on Robotic and Sensors Environments (ROSE)*. 25–30.

Kevin M. Lynch and Frank C. Park. 2017. *Modern Robotics.* Cambridge University Press.

Dushyant Mehta, Srinath Sridhar, Oleksandr Sotnychenko, Helge Rhodin, Mohammad Shafiei, Hans-Peter Seidel, Weipeng Xu, Dan Casas, and Christian Theobalt. 2017. VNect: Real-time 3D Human Pose Estimation with a Single RGB Camera. *ACM Transactions on Graphics* 36, 4, 1–14.

Shohin Mukherjee, Deepak Paramkusam, and Santosha K. Dwivedy. 2015. Inverse kinematics of a NAO humanoid robot using kinect to track and imitate human motion. In *Proceedings of the International Conference on Robotics, Automation, Control and Embedded Systems (RACE).* 1–7.

Atsushi Nakazawa, Shinichiro Nakaoka, Katsushi Ikeuchi, and Kazuhito Yokoi. 2002. Imitating human dance motions through motion structure analysis. In *Proceedings of the International Conference on Intelligent Robots and Systems (IROS)*, Vol. 3. 2539–2544.

Chrystopher L. Nehaniv and Kerstin Dautenhahn. 1998. Mapping between dissimilar bodies: Affordances and the algebraic foundations of imitation. In *Proceedings of the European Workshop on Learning Robots (EWLR-7).* 64–72.

Chrystopher L. Nehaniv and Kerstin Dautenhahn. 2007. *Imitation and Social Learning in Robots, Humans and Animals: Behavioural, Social and Communicative Dimensions.* Cambridge University Press.

Rui Ogawa, Sung Park, and Hiroyuki Umemuro. 2019. How Humans Develop Trust in Communication Robots: A Phased Model Based on Interpersonal Trust. In *Proceedings of the International Conference on Human-Robot Interaction (HRI).* 606–607.

Yongsheng Ou, Jianbing Hu, Zhiyang Wang, Yiqun Fu, Xinyu Wu, and Xiaoyun Li. 2015. A real-time human imitation system using kinect. *International Journal of Social Robotics* 7 (2015), 587–600.

Eunil Park, Hwayeon Kong, Hyeong-taek Lim, Jongsik Lee, Sangseok You, and Angel Pasqual del Pobil. 2011. The effect of robot's behavior vs. appearance on communication with humans. In *Proceedings of the International Conference on Human-Robot Interaction (HRI).* 219–220.

Luigi Penco, Nicola Scianca, Valerio Modugno, Leonardo Lanari, Giuseppe Oriolo, and Serena Ivaldi. 2019. A Multimode Teleoperation Framework for Humanoid Loco-Manipulation: An Application for the iCub Robot. *IEEE Robotics & Automation Magazine* 26, 4 (2019), 73–82.

Rifca Peters, Joost Broekens, and Mark A. Neerincx. 2017. Robots educate in style: The effect of con- text and non-verbal behaviour on children's perceptions of warmth and competence. In *Proceedings of the International Symposium on Robot and Human Interactive Communication (RO-MAN).* 449–455.

Hunter Rogers, Amro Khasawneh, Jeffery Bertrand, and Kapil Chalil Madathil. 2017. An investigation of the effect of latency on the operator's trust and performance for manual multi-robot teleoperated tasks. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting,* Vol. 61. 390–394.

Astrid M. Rosenthal-von der Pütten, Nicole C. Krämer, and Jonathan Herrmann. 2018. The effects of humanlike and robot-specific affective nonverbal behavior on perception, emotion, and behavior. *International Journal of Social Robotics* 10 (2018), 569–582.

Hamed Saeidi, John R Wagner, and Yue Wang. 2017. A mixed-initiative haptic teleoperation strategy for mobile robotic systems based on bidirectional computational trust analysis. *IEEE Transactions on Robotics* 33, 6 (2017), 1500–1507.

Maha Salem, Friederike Eyssel, Katharina Rohlfing, Stefan Kopp, and Frank Joublin. 2013. To err is human (-like): Effects of robot gesture on perceived anthropomorphism and likability. *International Journal of Social Robotics* 5 (2013), 313–323. https://doi.org/10.1007/s12369-013-0196-9

Wataru Sato and Sakiko Yoshikawa. 2007. Spontaneous facial mimicry in response to dynamic facial expressions. *Cognition* 104, 1 (2007), 1–18.

Stefan Schaal. 1999. Is imitation learning the route to humanoid robots? *Trends in Cognitive Sciences* 3, 6 (1999), 233–242.

Richard C. Schmidt, Samantha Morr, Paula Fitzpatrick, and Michael J. Richardson. 2012. Measuring the dynamics of interactional synchrony. *Journal of Nonverbal Behavior* 36 (2012), 263–279.

Amanda Sharkey and Noel Sharkey. 2020. We need to talk about deception in social robotics! *Ethics and Information Technology* (2020), 1–8.

Michihiro Shimada, Kazunori Yamauchi, Takashi Minato, Hiroshi Ishiguro, and Shoji Itakura. 2008. Studying the Influence of the Chameleon Effect on Humans using an Android. In *Proceedings of the International Conference on Intelligent Robots and Systems.* 767–772.

Jamie Shotton, Andrew Fitzgibbon, Mat Cook, Toby Sharp, Mark Finocchio, Richard Moore, Alex Kipman, and Andrew Blake. 2011. Real-time human pose recognition in parts from single depth images. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR).* 1297–1304.

Christopher Stanton, Anton Bogdanovych, and Edward Ratanasena. 2012. Teleoperation of a humanoid robot using full-body motion capture, example movements, and machine learning. In *Proceedings of the Australasian Conference on Robotics and Automation (ACRA).* 1–10.

Darja Stoeva, Helena A. Frijns, Margrit Gelautz, and Oliver Schürer. 2021. Analytical Solution of Pepper's Inverse Kinematics for a Pose Matching Imitation System. In *Proceedings of the International Conference on Robot & Human Interactive Communication (RO-MAN).* 167–174.

Nguyen T. V. Tuyen, Sungmoon Jeong, and Nak Y. Chong. 2018. Emotional Bodily Expressions for Culturally Competent Robots through Long Term Human-Robot Interaction. In *Proceedings of the International Conference on Intelligent Robots and Systems (IROS).* 2008–2013.

Linda N. Vallée, Sao M. Nguyen, Christophe Lohr, Ioannis Kanellos, and Olivier Asseu. 2020. How An Automated Gesture Imitation Game Can Improve Social Interactions With Teenagers With ASD. In *ICRA workshop on Social Robotics for Neurodevelopmental Disorders.*

Katsu Yamane and Akihiko Murai. 2016. A Comparative Study Between Humans and Humanoid Robots. *In Humanoid Robotics: A Reference,* Ambarish Goswami and Prahlad Vadakkepat (Eds.). 1–20. https://doi.org/10.1007/978-94-007-6046-2_7

Unai Zabala, Igor Rodriguez, José M. Martínez-Otzeta, and Elena Lazkano. 2020. Can a Social Robot Learn to Gesticulate Just by Observing Humans? In *Advances in Physical Agents II,* Luis M. Bergasa, Manuel Ocaña, Rafael Barea, Elena López-Guillén, and Pedro Revenga (Eds.). 137–150. https://doi.org/10.1007/978-3-030-62579-5_10

Liang Zhang, Zhihao Cheng, Yixin Gan, Guangming Zhu, Peiyi Shen, and Juan Song. 2016. Fast human whole body motion imitation algorithm for humanoid robots. In *Proceedings of the International Conference on Robotics and Biomimetics (ROBIO).* 1430–1435.

Ming Zhang, Jianxin Chen, Xin Wei, and Dezhou Zhang. 2018. Work chain-based inverse kinematics of robot to imitate human motion with Kinect. *ETRI Journal* 40, 4 (2018), 511–521.

Tehao Zhu, Qunfei Zhao, Weibing Wan, and Zeyang Xia. 2017. Robust regression-based motion perception for online imitation on humanoid robot. *International Journal of Social Robotics* 9 (2017), 705–725.

Fernando Zuher and Roseli Romero. 2012. Recognition of human motions for imitation and control of a humanoid robot. In *Proceedings of the Brazilian Robotics Symposium and Latin American Robotics Symposium (SBR-LARS).* 190–195.

Motion Planning for Human-Robot Collaboration

Florian Beck, Andreas Kugi

Abstract

In this work, we address motion planning for robots in human-robot collaboration. An overview of important properties of a motion planning algorithm in terms of safety and human comfort is provided. In terms of comfort, we emphasize fluency, legibility, and human-like motion. Furthermore, existing planning algorithms are reviewed and contrasted in terms of these desired properties. Based on this review of the literature, a receding horizon trajectory optimization approach is proposed, and its main features are highlighted.

Keywords

Motion Planning, Receding Horizon, Human-Robot Collaboration, Safety, Comfort

1 Introduction

In recent years, there has been an increase in demand for robots capable of working in the proximity of humans or even collaborate with them. Possible applications range from collaborative tasks in industry, such as load sharing tasks or joint assembly, to service robots in domestic environments. Because traditional safety measures such as fences are no longer appropriate for these applications, novel concepts are required to enable safe collaboration. Aside from safety, collaborative tasks give rise to additional requirements in task orchestration and adaptable robot behavior based on observations of the environment. Furthermore, human comfort during the interaction is critical in establishing the robot as a trustworthy collaborator.

These requirements are typically handled by different layers in the automation architecture. First, a cognitive decision layer coordinates tasks between the human and the robot. This layer gives explicit goals to a motion planning layer, which are then executed by an underlying controller layer.

In this work, we focus on the motion planning layer while explicitly considering the interface to a suitable controller for task execution. In the first step, an overview of the requirements with respect to safety and human comfort in human-robot collaboration (HRC) will be provided. Second, existing planning algorithms proposed in the literature will be shortly reviewed given these requirements.

Based on this analysis, open issues towards a flexible motion planning approach for HRC are identified. A receding horizon trajectory optimization planner is proposed as a contribution to resolving these issues. For this, we take advantage of the possibility to formulate the requirements for safety and comfort during the interaction as objective functions and constraints for trajectory optimization.

https://doi.org/10.34727/2022/isbn.978-3-85448-052-5_6

2 Collaborative Robots

Collaborative robotics applications require not only algorithmic solutions for the control, planning, and cognitive layers, but also suitable mechanical structures. In recent years, several collaborative robots, also referred to as cobots, have been developed and brought to market. Examples include robots by Universal Robots, the KUKA LBR iiwa, and Franka Emika's Panda. The latter two are based on technology developed at DLR [Hirzinger et al. 2002] focusing on lightweight, torque-controlled robots with elastic joints. The main advantage of the lightweight design, while maintaining a reasonable payload, is that it reduces the inertia of the robot links which directly contributes to reducing injuries upon impact. Another benefit of such lightweight collaborative robots is their ability to be mounted on mobile platforms, allowing for mobile manipulation. Furthermore, these robots feature seven degrees of freedom (DOF), which increases the manipulability through kinematic redundancy. The 7 DOF robot arms also mimic the human hand to some extent, allowing analogies in planning and control to be drawn between the human and the robot.

In our work, we use the KUKA LBR iiwa 14 R820 as a collaborative manipulator arm, which can also be mounted on DS Automotion's Sally, a differential drive mobile platform, shown in Fig. 1 as a reference platform.

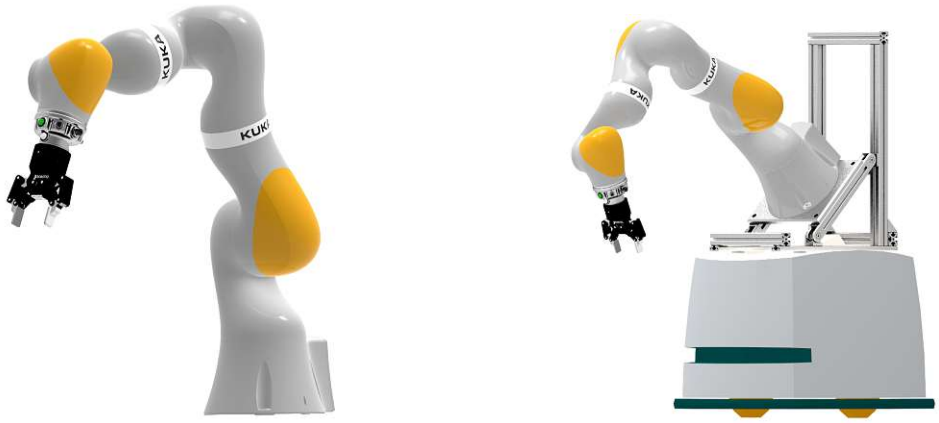


Figure 1 The KUKA LBR iiwa 14 R820 collaborative robot with a gripper (left) and KUKA LBR iiwa 14 R820 mounted on DS Automotion's SALLY, a differential drive mobile platform (right).

3 Motion Planning Requirements for HRC

Motion planning algorithms serve as a central component in the robot's automation architecture. In this section, the desired properties of motion planning algorithms with respect to human-robot collaboration, in particular safety and comfort, as well as existing approaches from the literature will be reviewed.

3.1 Safety

The most important criterion for enabling close collaboration between humans and robots is safety. In the context of industrial robots, safety is typically ensured by fencing or structural measures surrounding the robot, such that the robot moves only if no humans are in close proximity. This is of course incompatible with collaborative tasks. As a result, safety concepts are required to avoid collisions or to mitigate the consequences of impact in case of collisions. An overview of design criteria for safe human-robot interaction, both on the mechanical construction level and for the algorithm design, is given in [Alami et al. 2006]. In the following, we focus on algorithm design, assuming appropriate mechanical properties as described in Section 2.

In [Haddadin et al. 2017], the authors distinguish between two phases, namely pre-collision and post-collision. Motion planning is mainly concerned with the pre-collision phase. This means that collision-free trajectories must be planned while still meeting of task completion requirements. Typical collision avoidance approaches require a geometric representation of the robot's environment. Due to the computational complexity, objects are often approximated by convex shapes. Although this simplifies pairwise collision checks, the environment can still be non-convex if it contains multiple convex shapes. Using these convex representations, algorithms like the Gilbert-Johnson-Keerthi (GJK) algorithm [Gilbert et al. 1988; Cameron 1997] can be employed to check whether the robot is in collision with the environment in a given configuration. Another popular collision checking approach is the V-Clip Algorithm [Mirtich 1998]. The paper's performance comparison does not indicate a clear improvement, but rather depends on the specific application.

Collision checking is computationally expensive in motion planning in general because pairwise checks between obstacles and the robot, or parts of the robot, must be performed. Furthermore, depending on the planning algorithm this may have to be repeated several times. For safety, it is also important to consider that such collision checking approaches are only executed at discrete points in time. Collisions between two sampling points are theoretically possible depending on

the time discretization intervals. There are two solutions to this problem. First, sampling density can be increased. This, however, comes at significant computational costs due to the increased number of samples required for representing a movement. Second, collision detection can be extended to a continuous-time approach. Examples for continuous collision checking can be found in [Schulman et al. 2014] and [Merkt et al. 2019]. Such collision detection approaches are also applicable in dynamic environments. However, the environment geometry must depend on time. Hence, a motion model of objects in the environment is required to predict their movement.

For motion planning approaches, the post-collision phase must be considered in addition to the pre-collision phase. The post-collision is typically treated in the underlying control layer. Detection of collision and appropriate reaction strategies are proposed in [De Luca et al. 2006] and [Haddadin et al. 2008] through torque measurements in the joints. These strategies are typically combined with impedance control [Ott 2008], enabling a compliant robot behavior. Such compliant behavior is often desired in the Cartesian space of the robot end-effector. To use such control laws, the motion planning algorithm must provide sufficiently smooth trajectories, i.e. at least two times continuously differentiable. Furthermore, due to the presence of the inverse Jacobian in the control law, Cartesian impedance control requires singularity-free trajectories. In this regard, it is critical not only to avoid singular configurations during planning, but also to include a sufficiently large safety margin around the singularities. This is because, in the proximity of singularities, small velocities in the task space can still result in large velocities in the joint space.

3.2 Natural Motion and Comfort

In addition to functional aspects of a planner, such as reaching a goal, feasibility of the trajectory, and the adherence to safety aspects according to Section 3.1, human comfort must be taken into account when planning a robot's motion. In general, it is difficult to rigorously define robot motion that is comfortable for humans. It is highly dependent on how a human perceives the situation and can vary greatly depending on the individual. Furthermore, it may be dependent on the robot's capabilities and design. Studies in human-robot interaction (HRI) try to identify such properties. Furthermore, it is desirable to formalize such properties to an extent such that they can be considered during planning. This was accomplished for certain criteria, which will be discussed in the following. An overview of social aspects and psychological factors for safety in HRI can be found in [Lasota et al. 2017].

One of the most discussed aspects regarding comfort is proxemics [Hall 1963], i.e. the notion of distance between humans or a human and a robot, respectively, during certain interactions. The influence of a separation distance between a robot and a human was for example investigated in [Arai et al. 2010; Koay et al. 2006; Kulić and Croft 2007]. In a collaborative setting, distance is frequently constrained by the task at hand. Aside from the desired end-effector goal, there are often additional DOF that can be used to determine the pose or movement of the robot in space, depending on the specific task. For a mobile manipulator, this includes the positioning of the vehicle itself concerning the end-effector goal and the human.

An important concept with respect to comfort is legibility, as for example discussed in [Lichtenthäler et al. 2012] to increase the perceived safety. Legibility is a measure of how well the robot can convey its intent. In the motion planning context, this means that movement has to be planned such that ambiguity is reduced making goals easily inferable by a human. In some cases, this can be achieved by certain exaggeration of the movement, for example moving in a circular arc toward an object. Of course, this type of exaggeration is not always achievable, especially if several target objects are located close to each other. In such a scenario, it depends on other factors, e.g. if the human can infer where the robot is moving next. This cannot be solved using motion planning alone. An optimization-based formulation of legibility can be found in [Dragan et al. 2013], which also gives a comparison to the notion of predictability. Predictable motion is defined as predicting how a motion will look like if the goal has already been determined. As a result, the inference direction is reversed. In this regard, predictable motion can differ from legible motion. Predictability or legibility is preferred depending on the collaborative task at hand. For example, if the task consists of a fixed, sequential process, a human already knows what the robot's goals are, and predictability is more important than legibility. Legible motion, on the other hand, is preferred when the task is ambiguous.

In [Hoffman 2019], an overview of methods to evaluate fluency in HRC is given. They provide a definition and a model for assessing fluency. Fluent collaboration occurs when a human and a robot achieve a high level of coordination, resulting in precisely timed, efficient sequences of action. In a user study, they discovered that human idle time, i.e. the human waiting for the robot, as well as the functional delay of the robot, has a significant influence on subjective fluency. Longer human idle time is perceived as increasing fluency, which was indicated by feedback from participants who thought the robot did a better job. Increasing the functional delay, on the other hand, has a negative impact on the sense of fluency. This can be directly related to the robot's time to action following the completion of the human's turn during the collaboration. The requirement of short functional delays implies that fast planning and replanning are essential properties of motion

planning algorithms. An example of fluency for robot-human handovers is given in [Cakmak et al. 2011] considering the functional delay. They propose that conveying intent is a major factor in fluency. If the robot does not make its intentions to hand over an object clear, functional delays increase and the sense of fluency decreases during the interaction. This demonstrates that not only fast planning is required, but approach directions and timing must also be considered for comfortable interactions. Further examples of the importance of approach directions during handovers are given in [Koay et al. 2007] and [Sisbot and Alami 2012]. Human motion and action prediction are extremely useful for reducing such functional delays and increasing fluency. There is a substantial body of literature on human motion prediction in terms of long-term prediction, i.e. full reaching motions, see, e.g., [Luo et al. 2018], as well as short-term predictions obtained by tracking algorithms. Both are important for motion planning. Short-term predictions primarily improve the observations, resulting in more accurate estimates of the goals and dynamic obstacles in the environment. Long-term prediction, on the other hand, can be used to estimate human intentions and thus, influence fluency directly. Prediction combined with rapid replanning results in both reactive and anticipatory action [Hoffman and Breazeal 2007].

Depending on the mechanical structure of the robot, also human-like motions can be planned. Anthropomorphic robot arms, for example, such as the KUKA LBR iiwa, mimic the structure of a human arm with seven DOF. Optimal control theory was used to analyze human reaching motions in relation to the hand pose, see, e.g., [Flash and Hogan 1985] and [Todorov and Jordan 2002]. The results show that hand movement minimizes jerk, leading to smooth motions with bell-shaped velocity profiles. These findings provide explicit criteria that, in principle, can be applied to robotic motion planning. Maximizing the smoothness of the trajectories is somehow contradictory to minimizing the time, i.e. time optimality, which is commonly desired in industrial processes to maximize throughput. Fast robot movements, however, are perceived as less safe when interacting with humans [Arai et al. 2010]. This implies that the smoothness of robot motion is extremely important in HRC. Another important aspect is motion planning in the task space, i.e. Cartesian end-effector coordinates because most existing motion planning algorithms are designed in the joint space. In the case of a redundant robot, the nullspace motion must also be considered. The nullspace motion typically determines the robot's elbow movement, which is strongly dependent on the robot structure and can only be determined on a very limited basis by HRI.

4 Motion Planning Algorithms

In this section, we give an overview of existing motion planning algorithms in the literature while also assessing their capabilities with respect to the criteria identified in Section 3. Because of its importance in robot autonomy, motion planning has received a lot of attention in robotics research. The corresponding algorithms can be categorized into planning for static and dynamic environments. While we are primarily interested in real-time planning in dynamic environments, algorithms proposed for static environments are frequently used as the foundation for developing real-time capable methods for dynamic environments.

In static environments, sampling-based methods received a lot of attention. Their primary benefit is that obstacles do not need to be explicitly modeled in the configuration space. A collision detection module is instead used to determine whether or not a sample in configuration space is in collision. This greatly improves the planning efficiency [LaValle 2006]. Two important representatives of sampling-based algorithms are Probabilistic Roadmap (PRM) [Kavraki et al. 1996] and Rapidly-exploring Random Trees (RRT) [LaValle and J. 2001]. While PRM invests heavily in preprocessing to provide fast multi-query planning, RRTs are designed to be fast single-query planners. The basic RRT algorithm has probabilistic completeness, i.e. in the limit a path, if it exists, will be obtained with probability one. For a simplified version of PRM, this was proven as well [Kavraki et al. 1998]. Since their initial publication, several extensions were proposed to PRM and RRT motion planning. For our purpose, extensions toward optimal motion planning are the most relevant. Thus, for instance, the asymptotically optimal algorithms RRT* and PRM* were proposed in [Karaman and Frazzoli 2011]. Although sampling-based motion planners have several desirable properties, particularly probabilistic completeness, they frequently suffer from non-smooth trajectories, which require further post-processing. This ultimately increases the planning time. Furthermore, complex objectives and constraints lead to a high computational load. This can be a problem when formulating the objectives for comfort, as discussed in Section 3.2.

As a possible solution to these issues, trajectory optimization was proposed. Although, in general, trajectory optimization returns only locally optimal trajectories, it has been successfully applied to robotic motion planning. Trajectory optimization can be used to refine trajectories obtained from sampling-based planners, but it can also be used as a stand-alone algorithm. In [Ratliff et al. 2009; Zucker et al. 2013], an optimization-based planner called Covariant Hamiltonian Optimization for Motion Planning (CHOMP) was proposed. The objective function consists of two cost terms, an obstacle cost based on Euclidean distance fields and a smoothness cost that takes velocities and accelerations into account along the trajectory.

The trajectory is updated iteratively using covariant gradient descent. The update rule ensures that the trajectory remains smooth while decreasing the cost. The experiments demonstrate the algorithm's successful application to robotic manipulation. One significant drawback, which the authors also mention, is that due to the fixed discretization, only trajectories of a predefined length are considered.

In [Kalakrishnan et al. 2011], a stochastic optimization approach for motion planning called Stochastic Trajectory Optimization for Motion Planning (STOMP) is presented. The authors propose using a series of noisy trajectories that deviate slightly from the current candidate trajectory, and are then simulated to determine their costs. The candidate solution is updated based on these costs. One of the main advantages of this approach is that, because of derivative-free stochastic optimization, it can deal with general constraints for which gradients are not always available. This can be an advantage compared to CHOMP [Ratliff et al. 2009; Zucker et al. 2013] if desirable cost functions are not differentiable.

The method in [Schulman et al. 2014] is similar to CHOMP [Ratliff et al. 2009; Zucker et al. 2013], however, the authors make use of sequential convex optimization. In each iteration, a convex approximation of the nonlinear trajectory optimization problem is constructed. A trust region method is used to ensure that the approximation remains valid. In addition, infeasible constraints are converted to ℓ_1 penalties. A quadratic programming solver is used to solve the convex subproblem. For collision checking, GJK as mentioned in Section 3.1 is used. To ensure continuous-time safety, the collision checking procedure takes into account a swept-out volume, which is a polyhedral approximation of the free configuration space between two time steps. When compared to CHOMP [Ratliff et al. 2009; Zucker et al. 2013] and sampling-based planners implemented in the open motion planning library (OMPL) [Şucan et al. 2012] including RRT [LaValle and J. 2001], the experiments show a significant improvement in terms of speed, the problems that can be solved, and the quality of the resulting trajectories. Furthermore, this framework allows for inclusion of more complex cost functions, such as those related to human comfort.

Recently, a framework for guaranteed sequential trajectory optimization (GuSTO) [Bonalli et al. 2019] using sequential convex programming (SCP) was proposed. In contrast to TrajOpt [Schulman et al. 2014], which makes use of SCP as well, theoretical guarantees for convergence to at least a stationary point are given by the authors. Numerical simulations demonstrate that this approach provides more accurate results in less time compared to other state-of-the-art SCP-based planners.

To capture dynamic environments and real-time planning, several approaches can be found in the literature. Extensions to RRT planning include [Li and Shie

2002; Ferguson et al. 2006] and [Zucker et al. 2007]. In addition, [Svenstrup et al. 2010] use the RRT algorithm in combination with a dynamic potential field. The potential field takes into account the robot's position in the environment, its goal, and the humans moving in its vicinity. To account for changes in the environment the planner is implemented as a model predictive controller (MPC). To that end, only the first few steps of the planned trajectory are executed, while the planner calculates a new trajectory on-line. In [Sun et al. 2015], a similar RRT-based approach for high-frequency replanning was developed. A stochastic motion model of the robot is used. Several independent RRTs are executed in parallel to quickly find an optimal plan. The lowest cost plan is then chosen. While a single RRT will not find an optimal solution, it is proven that running several RRTs in parallel will asymptotically converge to an optimal plan. However, sampling-based planners for dynamic environments have the same drawbacks as their static counterparts.

In [Park et al. 2012], a similar concept using trajectory optimization is proposed. The motion of dynamic obstacles is taken into account by predicting their motion over a short-time horizon and computing a conservative local bound on their location and velocity. Based on this information, a constraint optimization problem is solved to compute a plan. Because dynamic object trajectories are only predicted for a short period of time, prediction uncertainty grows quickly. The planner is executed again in each time step, and only one step of the trajectory is executed before replanning.

The works [Ghazaei Ardakani et al. 2015, 2019] present an MPC approach for real-time point-to-point trajectory generation for a robot manipulator. A linear kinematic robot model is used, given by a double integrator system, where joint positions, velocities and accelerations serving as optimization variables. The final trajectories are generated using linear interpolation with a fixed sampling time. Because of the fixed sampling intervals and the goal constraint on the final step, it is assumed that the trajectory duration is sufficient to reach the goal while taking the robot's kinematic limits into account. The authors successively reduce the sampling period in the experiments, increasing the time resolution of the trajectory as the robot approaches the goal. The fixed sampling period, on the other hand, implies that the robot trajectory is initially quite coarse, which can be problematic in terms of constraint satisfaction, such as collision constraints for safety. Due to the convex formulation of the optimization problem, the authors report fast convergence of their algorithm. The convex formulation, on the other hand, significantly limits the available optimization criteria.

In contrast, in [Krämer et al. 2020] a different approach utilizing a cost-to-go-term was proposed replacing the requirement of a goal constraint. This allows for a fewer discretization points along the trajectory without sacrificing sampling density. This is especially important in terms of safety because high sampling

density reduces the likelihood of collision between trajectory samples while being computationally less expensive than continuous collision checking as, for example, done in [Schulman et al. 2014]. The results of [Krämer et al. 2020] show that achieving planning times below 100 ms per MPC iteration for pick-and-place tasks is feasible.

In [Agboh and Dogar 2018], an extension of STOMP [Kalakrishnan et al. 2011] to real-time replanning for grasping in cluttered environments is proposed. Initially, an open-loop trajectory is generated with numerous iterations to obtain a locally optimal solution. Starting with this initial trajectory, replanning is done with fewer iterations and with feedback from the current state. High-quality trajectories can be generated while maintaining fast planning times if the inital trajectory is a good initialization for replanning. The experiments show that the approach works well for grasping in cluttered environments that do not change too quickly. For moving targets or obstacles, the initialization is not a good approach since the trajectory can already be infeasible when the planner has finished.

A local receding horizon trajectory optimization given a global reference path in a difficult terrain is proposed in [Howard et al. 2010]. In [Toit and Burdick 2012], robot motion planning is formulated as a stochastic dynamic programming (SDP) problem. The authors explicitly address uncertainty rooted in the robot's environment. Because of the stochastic context, constraints are formulated as chance constraints [Toit and Burdick 2011], meaning that the constraint has to be fulfilled with a certain confidence. Given the complexity of the SDP problem, it is approximately solved using stochastic receding horizon control in the belief space. In dynamic uncertain environments, the stochastic approach provides more accurate models for planning. However, when compared to deterministic solutions, the additional computational effort is significant.

Recently, an MPC concept for autonomous guided vehicle motion planning was published [Mercy et al. 2018]. The authors use B-Spline trajectory parametrization to guarantee constraint satisfaction in the resulting nonlinear trajectory optimization problem. In contrast to [Toit and Burdick 2012], obstacles are modeled and predicted in a deterministic way facilitating a linear prediction model. The experiments show that dynamic obstacles in the environment can be safely avoided when combined with fast replanning.

MPC can also be used to plan and track a robot's trajectory at the same time. This has the advantage of not requiring the use of a trajectory following controller. Furthermore, the dynamic constraints of the entire system can be systematically considered allowing for more aggressive trajectories. The MPC framework CIAO [Schoels et al. 2020] is based on a novel convex inner approximation of the collision avoidance constraint. This enables the planning of kinodynamically fea-

sible collision-free trajectories in continuous time. A real-world experiment with a differential drive mobile robot demonstrates the unified trajectory optimization and tracking. Planning for multi-body robots, on the other hand, has yet to be demonstrated.

Simultaneous trajectory optimization and tracking was also applied to full dynamic models of robot manipulators, see, e.g., [Tassa et al. 2012] and drones, see, e.g, [Neunert et al. 2016]. Recently, Kleff et al. [Kleff et al. 2021] proposed an MPC approach based on differential dynamic programming (DDP) in real time on a collaborative robot. However, so far, MPC with full dynamics has only been solved for simplified problems, with additional objectives such as obstacle avoidance being neglected. As a result, for the currently available real-time hardware, approaches with separate trajectory planning and trajectory tracking control are typically used.

5 Receding Horizon Trajectory Optimization

In this section, we provide a brief overview of a receding horizon trajectory optimization approach for robot motion planning that takes into account the requirements from Section 3. In comparison to previous works discussed in Section 4, we explicitly take into account the combined requirements from Section 3, namely pre-collision and post-collision safety, legibility and smooth robot motion while enabling fluent interaction. We maintain compatibility with Cartesian impedance control by introducing computationally efficient singularity avoidance based on penalty functions. In addition, a novel via-point approach for receding horizon trajectory optimization is discussed providing a framework for planning legible and human-like motion with low computational overhead. Due to our emphasis on computational efficiency in the aforementioned features, fluent interactions can be ensured.

The planning approach considers robot manipulators under kinematic constraints. Note that robot dynamics are not considered in the planner. It is assumed that the underlying controller, i.e. a Cartesian compliance control scheme [Ott 2008], compensates for the nonlinear dynamics resulting in a remaining linear double integrator system. The motion planning problem is formulated as a trajec-

tory optimization in the form

$$\min_{\mathbf{u}_{0|n},\dots,\mathbf{u}_{N-1|n}} \sum_{k=0}^{N-1} l(\mathbf{x}_{k|n}, \mathbf{u}_{k|n}) \tag{1a}$$

$$\text{s.t.} \quad \mathbf{x}_{k+1|n} = \mathbf{\Phi}\mathbf{x}_{k|n} + \mathbf{\Gamma}\mathbf{u}_{k|n} \tag{1b}$$

$$\mathbf{x}_{0|n} = \mathbf{x}_{1|n-1},\ \mathbf{u}_{0|n} = \mathbf{u}_{1|n-1} \tag{1c}$$

$$\underline{\mathbf{x}} \leq \mathbf{x}_{k|n} \leq \overline{\mathbf{x}}, \quad k = 0, \dots, N-1 \tag{1d}$$

$$\underline{\mathbf{u}} \leq \mathbf{u}_{k|n} \leq \overline{\mathbf{u}}, \quad k = 0, \dots, N-1 \tag{1e}$$

for the time steps $k = 0, \dots, N-1$, with fixed sampling time T_s. The optimization problem (1) is solved at every sampling instant n for the finite planning horizon NT_s. Only the first step of the optimal control input is applied to the system until the next sampling instant $n+1$. The optimization problem is then solved again, now starting one sampling time T_s ahead and therefore predicting one step further into the future. Hence, the planning horizon is said to be receding. Eq. (1a) describes a general objective function to be minimized for the planning horizon nT_s to $(n+N-1)T_s$ depending on the robot's state $\mathbf{x}_{k|n}$ and the input $\mathbf{u}_{k|n}$ at the time $(n+k)T_s$, $k = 0, \dots, N-1$. The objective function includes a cost term that represents the distance to the goal such that the robot moves toward this goal. Additional cost terms can be added depending on the specific application, which will be discussed in greater detail in the remainder of this section. The resulting linear system of the robot dynamics is an equality constraint defined by Eq. (1b). The planner is initialized from the previously calculated trajectory through Eq. (1c). State and input constraints, specifically addressing joint limits, velocity limits, and higher derivatives, if necessary, are considered in Eq. (1e) where $\underline{\mathbf{x}}$, $\underline{\mathbf{u}}$, and $\overline{\mathbf{x}}$, $\overline{\mathbf{u}}$ denote lower and upper bounds, respectively.

The receding horizon trajectory optimization shares advantages of the trajectory optimization approach over sampling based algorithms as stated in Section 4. This is particularly relevant to the flexibility of objective functions and constraints in modelling desired properties in human-robot interactions and ensuring smooth trajectories. Furthermore, we use a cost-to-go term for reaching the goal in combination with fixed sampling times similar to what is done in [Krämer et al. 2020]. This allows for fast planning while still maintaining tightly sampled trajectories. Fast planning times are essential to reduce the robot's functional delay, enabling fluent interactions. Nonetheless, safety cannot be sacrificed for the sake of fast planning times. As mentioned in Section 3.1, safety violations can in principle happen between the discrete time steps of the trajectory optimization. Due to the computational effort, we do not consider continuous-time collision checking but instead, rely on small sampling times T_s. This property is in contrast to previous approaches in the literature. For example, [Ghazaei Ardakani et al. 2015]

and [Mercy et al. 2018] demand that the final point in the planning horizon already reaches the goal. This requires either a fixed duration of the trajectory, i.e. independent of the distance to the goal or the introduction of the duration as an additional optimization variable, increasing the complexity and thus the computational effort of the problem.

Collision checking for receding horizon trajectory optimization can be performed using well-known approaches from the literature. However, the gradient of the objective function and constraints can be provided to improve the optimization algorithm's convergence behavior. To this end, we use a smooth distance approximation as introduced in [Vu et al. 2020]. We extend the formulation from rectangular boxes and points to spheres as basic obstacle shapes to allow for more complex environments. Because collision checking is frequently the bottleneck, limiting to simple shapes results in faster optimization times. In the context of collision checking, the receding horizon framework also enables planning in dynamic environments. New information about objects in the environment can be incorporated due to the constant replanning. Furthermore, by including object states in the dynamic constraints (1b), predictions of object movements in the planning horizon can be taken into account.

In addition to safety considerations in the pre-collision phase, we also address compatibility with Cartesian impedance control [Ott 2008] to enable post-collision safety. The Cartesian impedance controller requires trajectories to be at least two times continuously differentiable and singularity free. In the proposed approach, sufficient smoothness is guaranteed by the equality constraints (1b) and (1c). Note that, in general, the sampling times of the trajectories are significantly lower than those required for the execution of the controller. As a result, for the controller, the trajectories must be interpolated and resampled. The optimization framework provides several ways to make sure that the planned trajectories are free of singularities. As mentioned in Section 3.1, a safety margin around singularities has to be taken into account. A distance to a singular configuration can in general be defined by the so-called manipulability measure [Yoshikawa 1985]. Alternatively, if a robot's singular configurations are known, direct distance measures in the configuration space can be used. The safety margin can be formulated as an inequality constraint ensuring a minimum distance to singular configurations or by demanding a minimum amount of manipulability. In view of the computational costs, the singularity avoidance is realized by a penalty function which is added to the objective function $l(\mathbf{x}_{k|n}, \mathbf{u}_{k|n})$. Note that, in principle, this does not guarantee singularity-free motions due to competing cost terms, however, such a situation can be avoided by selecting a sufficiently large weight for the penalty function.

Besides safety, we explicitly address comfort discussed in Section 3.2 within the receding horizon trajectory optimization framework. First, we consider human-like movement, as for the example investigated in [Flash and Hogan 1985]. Human movement of the hand is regarded as minimizing jerk there. This corresponds to minimizing jerk along an end-effector trajectory in the task space in the robotic applications under consideration. This formulation can be easily incorporated into the trajectory optimization problem, however, it would require planning in the task space. Direct planning in the task space makes the consideration of the joint limit constraints more involved. Because of the nonlinear relationship between the joint and the task space, enforcing smoothness in the joint space is computationally more efficient but does not always result in human-like movement in the task space. Planning in the joint space, but formulating the cost-to-go term in the task space based on the forward kinematics is another option, again at the cost of higher computational effort. Thus, we propose to approximate a cost-to-go term in the task space by placing via-points in the task space along the planned trajectories. The approximation is more or less coarse and computationally expensive depending on the number of via-points. Such intermediate goals are also important for a variety of other robotic tasks. Grasping for example requires the gripper to be aligned with an object in a so-called pre-grasp pose before the grasp point is reached. Furthermore, via-points can also aid the establishment of comfortable interactions by ensuring appropriate approach directions and end-effector orientations. In addition, intermediate goals can help to disambiguate goals resulting in legible robot trajectories. Again, this is a computationally efficient approximation compared to what is done, e.g. in [Dragan et al. 2013] to achieve legible robot motion. In previous works, see, e.g., [Schulman et al. 2014; Ghazaei Ardakani et al. 2015], a common approach for intermediate goals was to constrain points along the trajectory to via-points. This requires predetermined timings for the via-points along the trajectory. Furthermore, for a receding planning horizon, this approach is not feasible because the via-point may not be reachable within the horizon.

In contrast, we propose to formulate the optimization problem in such a way that only the relative timing between via-points, i.e. a sequence of via-points, is considered, rather than the exact timing along the trajectory. To that end, we introduce a parametrized reference path that linearly interpolates from the starting configuration through the via-points to the goal. The path parameter's dynamics are added to the optimization problem to represent the progress along the path. In contrast to classical path-following control, see, e.g., [Böck and Kugi 2014, 2016; Faulwasser et al. 2017], we are not interested in precisely following the path, but only in accurately passing through the via-points. Therefore, the cost weights are adjusted so that progress along the path is favored over precise tracking between via-points. To pass the via-points exactly, a path progress dependent constraint

is introduced. Instead of being specified in advance, the optimizer determines the timings and velocities through the via-points in this formulation.

6 Conclusions

In this work, we outlined the requirements for motion planning algorithms in collaborative human-robot tasks. In addition to physical properties of collaborative robots, a brief overview of algorithmic safety measures for planning algorithms, particularly collision checking, was provided. Although safety is the topmost priority, it is not the only requirement for planning in human-robot collaborative tasks. In this context, properties related to comfort including proximity, legibility, fluency, and human-likeness of robot motion were discussed. Furthermore, the state-of-the-art motion planning algorithms were evaluated concerning these requirements. Finally, a motion planning framework based on a receding horizon optimization approach was outlined. This method enables the flexible specification of control objectives and the systematic incorporation of constraints to easily adjust the desired properties for HRC.

Bibliography

Wisdom C. Agboh and Mehmet R. Dogar. 2018. Real-Time Online Re-Planning for Grasping Under Clutter and Uncertainty. *IEEE-RAS International Conference on Humanoid Robots*, 1–8. https://doi.org/10.1109/HUMANOIDS.2018.8625041

R. Alami, A. Albu-Schaeffer, A. Bicchi, R. Bischoff, R. Chatila, A. De Luca, A. De Santis, G. Giralt, J. Guiochet, G. Hirzinger, F. Ingrand, V. Lippiello, R. Mattone, D. Powell, S. Sen, B. Siciliano, G. Tonietti, and L. Villani. 2006. Safe and dependable physical human-robot interaction in anthropic domains: State of the art and challenges. *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 1–16. https://doi.org/10.1109/IROS.2006.6936985

T. Arai, R. Kato, and M. Fujita. 2010. Assessment of operator stress induced by robot collaboration in assembly. *CIRP Annals* 59, 1 (2010), 5–8. https://doi.org/10.1016/j.cirp.2010.03.043

Martin Böck and Andreas Kugi. 2014. Real-time Nonlinear Model Predictive Path-Following Control of a Laboratory Tower Crane. *IEEE Transactions on Control Systems Technology* 22, 4 (2014), 1461–1473. https://doi.org/10.1109/TCST.2013.2280464

Martin Böck and Andreas Kugi. 2016. Constrained model predictive manifold stabilization based on transverse normal forms. *Automatica* 74 (2016), 315–326. https://doi.org/10.1016/j.automatica.2016.07.046

Riccardo Bonalli, Abhishek Cauligi, Andrew Bylard, and Marco Pavone. 2019. GuSTO: Guaranteed Sequential Trajectory optimization via Sequential Convex Programming.

International Conference on Robotics and Automation, 6741–6747. https://doi.org/10.1109/ICRA.2019.8794205

Maya Cakmak, Siddhartha S. Srinivasa, Min Kyung Lee, Sara Kiesler, and Jodi Forlizzi. 2011. Using spatial and temporal contrast for fluent robot-human hand-overs. *ACM/IEEE International Conference on Human-Robot Interaction,* 489–496. https://doi.org/10.1145/1957656.1957823

Stephen Cameron. 1997. Enhancing GJK: computing minimum and penetration distances between convex polyhedra. *International Conference on Robotics and Automation*, 3112–3117. https://doi. org/10.1109/ROBOT.1997.606761

Alessandro De Luca, Alin Albu-Schaffer, Sami Haddadin, and Gerd Hirzinger. 2006. Collision Detection and Safe Reaction with the DLR-III Lightweight Manipulator Arm. *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 1623–1630. https://doi.org/10.1109/IROS.2006. 282053

Anca D Dragan, Kenton C T Lee, and Siddhartha S Srinivasa. 2013. Legibility and predictability of robot motion. *ACM/IEEE International Conference on Human-Robot Interaction,* 301–308. https: //doi.org/10.1109/HRI.2013.6483603

Timm Faulwasser, Tobias Weber, Pablo Zometa, and Rolf Findeisen. 2017. Implementation of Nonlinear Model Predictive Path-Following Control for an Industrial Robot. *IEEE Transactions on Control Systems Technology* 25,4(2017),1505–1511. https://doi.org/10.1109/TCST.2016.2601624

Dave Ferguson, Nidhi Kalra, and Stentz Anthony. 2006. Replanning with RRTs. *IEEE International Conference on Robotics and Automation*,1243–1248. https://doi.org/10.1109/ROBOT.2006. 1641879

Tamar Flash and Neville Hogan. 1985. The coordination of arm movements: an experimentally confirmed mathematical model. *Journal of Neuroscience* 5, 7 (1985), 1688–1703. https://doi.org/10. 1523/JNEUROSCI.05- 07- 01688.1985

M Mahdi Ghazaei Ardakani, Björn Olofsson, Anders Robertsson, and Rolf Johansson. 2015. Realtime trajectory generation using model predictive control. *IEEE International Conference on Automation Science and Engineering*,942–948. https://doi.org/10.1109/CoASE.2015.7294220

M Mahdi Ghazaei Ardakani, Björn Olofsson, Anders Robertsson, and Rolf Johansson. 2019. Model Predictive Control for Real-Time Point-to-Point Trajectory Generation. *IEEE Transactions on Automation Science and Engineering* 16, 2 (2019), 972–983. https://doi.org/10.1109/TASE.2018.2882764

Elmer G Gilbert, Daniel W Johnson, and S Sathiya Keerthi. 1988. A fast procedure for computing the distance between complex objects in three-dimensional space. *IEEE Journal on Robotics and Automation* 4, 2 (1988),193–203. https://doi.org/10.1109/56.2083

Sami Haddadin, Alin Albu-Schaffer, Alessandro De Luca, and Gerd Hirzinger. 2008. Collision Detection and Reaction: A Contribution to Safe Physical Human-Robot Interaction. *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 3356–3363. https://doi.org/10.1109/IROS.2008.4650764

Sami Haddadin, Alessandro De Luca, and Alin Albu-Schäffer. 2017. Robot Collisions: A Survey on Detection, Isolation, and Identification. *IEEE Transactions on Robotics* 33, 6 (2017), 1292–1312. https://doi.org/10.1109/TRO.2017.2723903

Edward T Hall. 1963. A System for the Notation of Proxemic Behavior. *American Anthropologist* 65, 5(1963),1003–1026. https://doi.org/10.1525/aa.1963.65.5.02a00020

Gerhard Hirzinger, Norbert Sporer, Alin Albu-Schaffer, M. Hahnle, Rainer Krenn, A. Pascucci, and Manfred Schedl. 2002. DLR's torque-controlled light weight robot III-are we reaching the technological limits now? *IEEE International Conference on Robotics and Automation*,1710–1716. https://doi.org/10.1109/ROBOT.2002.1014788

Guy Hoffman. 2019. Evaluating Fluency in Human–Robot Collaboration. *IEEE Transactions on Human-Machine Systems* 49, 3 (2019), 209–218. https://doi.org/10.1109/THMS.2019.2904558

Guy Hoffman and Cynthia Breazeal. 2007. Cost-Based Anticipatory Action Selection for Human–Robot Fluency. *IEEE Transactions on Robotics* 23, 5 (2007), 952–961. https://doi.org/10.1109/TRO.2007.907483

Thomas Howard, Colin Green, and Alonzo Kelly. 2010. Receding Horizon Model-Predictive Control for Mobile Robot Navigation of Intricate Paths. *Field and Service Robotics*, 69–78. https://doi.org/10.1007/978-3-642-13408-1_7

Mrinal Kalakrishnan, Sachin Chitta, Evangelos Theodorou, Peter Pastor, and Stefan Schaal. 2011. STOMP: Stochastic trajectory optimization for motion planning. *IEEE International Conference on Robotics and Automation*, 4569–4574. https://doi.org/10.1109/ICRA.2011.5980280

Sertac Karaman and Emilio Frazzoli. 2011. Sampling-based algorithms for optimal motion planning. *The International Journal of Robotics Research* 30, 7 (2011), 846–894. https://doi.org/10.1177/0278364911406761

Lydia E Kavraki, Mihail N Kolountzakis, and Jean-Claude Latombe. 1998. Analysis of probabilistic roadmaps for path planning. *IEEE Transactions on Robotics and Automation* 14, 1 (1998), 166–171. https://doi.org/10.1109/70.660866

Lydia E Kavraki, Petr Svestka, Jean-Claude Latombe, and Mark H Overmars. 1996. Probabilistic roadmaps for path planning in high-dimensional configuration spaces. *IEEE Transactions on Robotics and Automation* 12, 4 (1996), 566–580. https://doi.org/10.1109/70.508439

Sebastien Kleff, Avadesh Meduri, Rohan Budhiraja, Nicolas Mansard, and Ludovic Righetti. 2021. High-Frequency Nonlinear Model Predictive Control of a Manipulator. *IEEE International Conference on Robotics and Automation*, 7330–7336. https://doi.org/10.1109/ICRA48506.2021.9560990

Kheng L Koay, Kerstin Dautenhahn, Sarah N Woods, and Michael L Walters. 2006. Empirical Results from Using a Comfort Level Device in Human-Robot Interaction Studies. *ACM SIGCHI/SIGART Conference on Human-Robot Interaction*,194–201. https://doi.org/10.1145/1121241.1121276

Kheng L Koay, Emrah A Sisbot, Dag S Syrdal, Mick L Walters, Kerstin Dautenhahn, and Rachid Alami. 2007. Exploratory Study of a Robot Approaching a Person in the Context of Handing Over an Object. *AAAI Spring Symposium: Multidisciplinary Collaboration for Socially Assistive Robotics*, 18–24.

Maximilian Krämer, Christoph Rösmann, Frank Hoffmann, and Torsten Bertram. 2020. Model predictive control of a collaborative manipulator considering dynamic obstacles. *Optimal Control Applications and Methods* 41, 4 (2020), 1211–1232. https://doi.org/10.1002/oca.2599

Dana Kulić and Elizabeth Croft. 2007. Physiological and subjective responses to articulated robot motion. *Robotica* 25, 1 (2007), 13–27. https://doi.org/10.1017/S0263574706002955

Przemyslaw A Lasota, Terrence Fong, and Julie A Shah. 2017. A Survey of Methods for Safe Human-Robot Interaction. *Foundations and Trends in Robotics* 5, 4 (2017), 261–349. https://doi.org/10.1561/2300000052

Steven M LaValle. 2006. *Planning Algorithms.* Cambridge University Press, Cambridge, U.K.

Steven M LaValle and James J Kuffner 2001. Rapidly-Exploring Random Trees: Progress and Prospects. In *Algorithmic and Computational Robotics*, Bruce Donald, Kevin Lynch, Daniela Rus (Hrsg.). A K Peters/CRC Press, NewYork. https://doi.org/10.1201/9781439864135

Tsai-Yen Li and Yang-Chuan Shie. 2002. An incremental learning approach to motion planning with roadmap management. *Proceedings 2002 IEEE International Conference on Robotics and Automation,* 3411–3416. https://doi.org/10.1109/ROBOT.2002.1014238

Christina Lichtenthäler, Tamara Lorenzy, and Alexandra Kirsch. 2012. Influence of legibility on perceived safety in a virtual human-robot path crossing task. *International Symposium on Robot and Human Interactive Communication*, 676–681. https://doi.org/10.1109/ROMAN.2012.6343829

Ruikun Luo, Rafi Hayne, and Dmitry Berenson. 2018. Unsupervised early prediction of human reaching for human–robot collaboration in shared workspaces. *Autonomous Robots* 42, 3 (2018), 631–648. https://doi.org/10.1007/s10514-017-9655-8

Tim Mercy, Ruben Van Parys, and Goele Pipeleers. 2018. Spline-Based Motion Planning for Autonomous Guided Vehicles in a Dynamic Environment. *IEEE Transactions on Control Systems Technology* 26, 6 (2018), 2182–2189. https://doi.org/10.1109/TCST.2017.2739706

Wolfgang Merkt, Vladimir Ivan, and Sethu Vijayakumar. 2019. Continuous-Time Collision Avoidance for Trajectory Optimization in Dynamic Environments. *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 7248–7255. https://doi.org/10.1109/IROS40897.2019.8967641

Brian Mirtich. 1998. V-Clip: Fast and Robust Polyhedral Collision Detection. *ACM Transactions on Graphics* 17, 3 (1998), 177–208. https://doi.org/10.1145/285857.285860

Michael Neunert, Cédric de Crousaz, Fadri Furrer, Mina Kamel, Farbod Farshidian, Roland Siegwart, and Jonas Buchli. 2016. Fast nonlinear Model Predictive Control for unified trajectory optimization and tracking. *IEEE International Conference on Robotics and Automation*, 1398–1404. https://doi.org/10.1109/ICRA.2016.7487274

Christian Ott. 2008. *Cartesian Impedance Control of Redundant and Flexible-Joint Robots.* Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-69255-3

Chonhyon Park, Jia Pan, and Dinesh Manocha. 2012. ITOMP: Incremental Trajectory Optimization for Real-Time Replanning in Dynamic Environments. *International Conference on Automated Planning and Scheduling*, 207–215.

Nathan Ratliff, Matt Zucker, J. Andrew Bagnell, and Siddhartha Srinivasa. 2009. CHOMP: Gradient optimization techniques for efficient motion planning. *IEEE International Conference on Robotics and Automation,* 489–494. https://doi.org/10.1109/ROBOT.2009.5152817

Tobias Schoels, Luigi Palmieri, Kai O. Arras, and Moritz Diehl. 2020. An NMPC Approach using Convex Inner Approximations for Online Motion Planning with Guaranteed Collision Avoidance. *IEEE International Conference on Robotics and Automation*, 3574–3580. https://doi.org/10.1109/ICRA40945.2020.9197206

John Schulman, Yan Duan, Jonathan Ho, Alex Lee, Ibrahim Awwal, Henry Bradlow, Jia Pan, Sachin Patil, Ken Goldberg, and Pieter Abbeel. 2014. Motion planning with sequential convex optimization and convex collision checking. *The International Journal of Robotics Research* 33, 9 (2014), 1251– 1270. https://doi.org/10.1177/0278364914528132

Emrah A Sisbot and Rachid Alami. 2012. A Human-Aware Manipulation Planner. *IEEE Transactions on Robotics* 28, 5 (2012), 1045–1057. https://doi.org/10.1109/TRO.2012.2196303

Ioan A Şucan, Mark Moll, and Lydia E Kavraki. 2012. The Open Motion Planning Library. *IEEE Robotics & Automation Magazine* 19, 4 (2012), 72–82. https://doi.org/10.1109/MRA.2012.2205651

Wen Sun, Sachin Patil, and Ron Alterovitz. 2015. High-Frequency Replanning Under Uncertainty Using Parallel Sampling-Based Motion Planning. *IEEE Transactions on Robotics* 31, 1 (2015), 104–116. https://doi.org/10.1109/TRO.2014.2380273

Mikael Svenstrup, Thomas Bak, and Hans J Andersen. 2010. Trajectory planning for robots in dynamic human environments. *IEEE/RSJ International Conference on Intelligent Robots and Systems,* 4293–4298. https://doi.org/10.1109/IROS.2010.5651531

Yuval Tassa, Tom Erez, and Emanuel Todorov. 2012. Synthesis and stabilization of complex behaviors through online trajectory optimization. *IEEE/RSJ International Conference on Intelligent Robots and Systems,* 4906–4913. https://doi.org/10.1109/IROS.2012.6386025

Emanuel Todorov and Michael I Jordan. 2002. Optimal feedback control as a theory of motor coordination. *Nature Neuroscience* 5, 11 (2002), 1226–1235. https://doi.org/10.1038/nn963

Noel E Du Toit and Joel W Burdick. 2011. Probabilistic Collision Checking With Chance Constraints. *IEEE Transactions on Robotics* 27, 4 (2011), 809–815. https://doi.org/10.1109/TRO.2011.2116190

Noel E Du Toit and Joel W Burdick. 2012. Robot Motion Planning in Dynamic, Uncertain Environments. *IEEE Transactions on Robotics* 28, 1 (2012), 101–115. https://doi.org/10.1109/TRO.2011.2166435

Minh N Vu, Patrik Zips, Amadeus Lobe, Florian Beck, Wolfgang Kemmetmüller, and Andreas Kugi. 2020. Fast motion planning for a laboratory 3D gantry crane in the presence of obstacles. *IFAC-PapersOnLine* 53, 2 (2020), 9508–9514. https://doi.org/10.1016/j.ifacol.2020.12.2427

Tsuneo Yoshikawa. 1985. Manipulability and redundancy control of robotic mechanisms. *IEEE International Conference on Robotics and Automation*, 1004–1009. https://doi.org/10.1109/ROBOT. 1985.1087283

Matt Zucker, James Kuffner, and Michael Branicky. 2007. Multipartite RRTs for Rapid Replanning in Dynamic Environments. *IEEE International Conference on Robotics and Automation,* 1603–1609. https://doi.org/10.1109/ROBOT.2007.363553

Matthew Zucker, Nathan Ratliff, Anca Dragan, Mihail Pivtoraiko, Matthew Klingensmith, Christopher Dellin, James Andrew Bagnell, and Siddhartha Srinivasa. 2013. CHOMP: Covariant Hamiltonian Optimization for Motion Planning. *International Journal of Robotics Research* 32 (2013), 1164–1193. https://doi.org/10.1177/0278364913488805

I See What You Did There: Towards a Gaze Mechanism for Joint Actions in Human-Robot Interaction

Michael Koller, Astrid Weiss, Markus Vincze

Abstract

We imagine that service robots must collaborate with humans in physical object manipulation tasks to be of assistance in everyday scenarios, such as setting a table. This collaboration requires the capability of joint attention to smoothly accomplish a shared goal. One special modality for joint attention is the gaze behavior of an actor. Herein, we discuss the human gaze in physical tasks and its underlying cognitive mechanisms, a novel probabilistic robotic gaze controller in object-centred collaborative physical tasks, and its inclusion in a well-known joint action human-robot interaction (HRI) benchmark. First, we discuss human gaze behavior as an important modality for signaling, detecting, and monitoring joint attention processes. This is followed by an overview of joint attention implementations in HRI and commonly used artificial intelligence methods for planning and plan recognition. These methods are used to mimic qualities of different components in psychological joint attention models in humans. In object manipulation tasks, the gaze behavior is not only used to gather information about the environment, but also has a communicative role, as the gaze direction can be interpreted by the interaction partner. The intended actions and beliefs about the current world state are communicated through the gaze. We argue that robotic gaze behavior, which humans easily interpret, will improve the interaction capability of a social robot. We investigate this claim in an already established HRI joint action benchmark scenario of collaboratively building a tower out of different blocks. To this end, we propose a stochastic gaze controller for joint action tasks and present results of a pilot study.

Keywords

human-robot interaction, joint attention, joint action, gaze, eye-tracking

1 Introduction

Think of a situation where you have to coordinate with another person in a physical task at hand. Let us say that you and a friend attempt to move a sofa up a staircase. Both of you have the same goal, namely, to bring the sofa up into another apartment, and the sofa would be too heavy for either one of you, to attempt to do so alone. Hence, each of you grabs one end of it. It is also clear to you that your actions influence each other, such that you must monitor and react to each other. Similarly, you can signal to your friend how you imagine to squeeze the sofa around the tight corner up ahead. You probably will not verbalize each and every intention, but you just push the sofa in one direction more than strictly necessary to signal a direction, or you catch the gaze of your friend by intently looking into their eyes, and then gaze into a direction you intend to go. A short nod on their side could signal that they understood. Both of you proceed just for a few seconds with the now shared and agreed upon plan, until you have to check in with your friend to coordinate again.

Collaboration is highly necessary and not overly mentally taxing for humans. Nevertheless, when paying close attention to these collaborative processes that

https://doi.org/10.34727/2022/isbn.978-3-85448-052-5_7

occur almost automatically, it seems that there are numerous different components on different levels of abstraction at work. For example, how do we notice the focused attention of others? Which mental processes let us adapt and align our plans? How do we infer the plans of others? How do we make sure that the other person is really on the same page as us? How do we choose which kind of signal to use for which kind of information? How do we draw the attention of others and signal attention on our part? One must consider all these questions when implementing the capability of human-robot collaboration on a social robot.

In this chapter, we first contribute a discussion of results in psychology related to this topic. Specifically, we review research on joint attention [Baron-Cohen 1994; Mundy and Newell 2007] and theory of mind [Baron-Cohen 1997] with a focus on the human gaze in physical tasks. These are important building blocks generally required for the success of collaborative tasks in human-human interaction (HHI). First, we properly differentiate the two terms and observe how theory of mind builds on joint attention. Then, we focus on joint attention in the robotic context. We contribute a review of different approaches employed by roboticists to provide robots with joint attention capability or at least a technically feasible equivalent. Finally, we propose a novel probabilistic robotic gaze controller for a joint action benchmark between the human and robot proposed by Clodic et al. [2017], based on building a tower out of various wooden blocks. For object-centered collaborative physical tasks, this represents an approach to generate realistic, intuitive, and interpretable gaze behavior. We report the initial results of a pilot study and discuss how to include it into the joint action benchmark. Our contribution extends a stochastic gaze controller for static scenarios to dynamic ones.

2 Joint Attention in Psychology

Joint attention has been studied since the 1970ies [Scaife and Bruner 1975]. Research on joint attention in psychology yielded structural and procedural models, as well as analyses whose cues are used to signal the state of joint attention between humans. If we intend to have service robots in the future that share environments with human beings and provide help in everyday physical tasks, they must be endowed with the ability to engage in joint attention [Krämer et al. 2011] in a similar way as two humans.

Joint attention is the process of sharing one's attention with another person, using social cues for coordination. The coordination effort focuses on a third object, event, or stimulus [Akhtar and Gernsbacher 2007]. One of the earliest reports of joint attention appeared 1975 in an article by Scaife and Bruner [1975] and studied the gaze following ability in infants. The experiment showed that only 30% of

two to four month old children engage in gaze following, whereas from the age of eleven months every infant is able to do so. To this day, a significant amount of research is conducted on joint attention in child development.

How can we achieve something functionally similar to human joint attention in *Social Robotics*? First, we consider some results of cognitive and social psychology to better understand how joint attention empowers humans. Furthermore, we consider the components constituting joint attention and how it is embedded in the broader coordination process.

2.1 On Theory of Mind and Modeling Joint Attention

One insightful approach is to recognize joint attention as a necessary building block for the more high-level mental capability of Theory of Mind (ToM). Tomasello [1995] describe joint attention and ToM as relevant in the field of social cognition, as they are concepts explaining how humans process information about other humans in social situations. Children at the end of their second year of life already possess the following capabilities: "1) They understand other persons in terms of their intentions. 2) They understand that others have intentions that may differ from their own. 3) They understand that others have intentions that may not match with the current state of affairs (accidents and unfulfilled intentions)." [Tomasello 1995, p. 105]

The term "theory of mind" was coined by Premack and Woodruff [1978] and comprises several mental capabilities that develop later in children, around the ages of three to four. It allows them to represent more complex mental states than intentions, namely: "1) They understand other persons in terms of their thoughts and beliefs. 2) They understand that others have thoughts and beliefs that may differ from their own. 3) They understand that others have thoughts and beliefs that may not match with the current state of affairs (false beliefs)."[Tomasello 1995, p. 104] [1]

Baron-Cohen [1994, 1997] claimed a structural relationship between the separate mental modules of joint attention and ToM. In fact, they claimed that the human ability they call "mind-reading" requires at least four components that build on each other. Mind-reading is defined in the sense that humans can often infer the thoughts, beliefs, plans, and emotional states of other people they observe or think about, in short, reason about "mental things."

1 Although the term *joint attention* originated in developmental psychology, other approaches in psychology also provided results on the topic, some of which is covered in the following subsections. In these, adults who exhibit a fully developed joint attention capability are the subject of the study. As our robot model is also not developmentally inspired, we do not focus on child development for the remainder of this chapter.

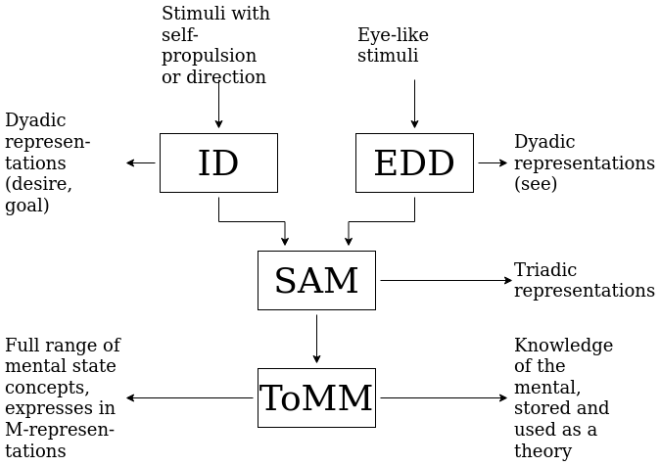


Figure 1 Mind-reading system, adapted from Baron-Cohen [1994].

The four component system consists of the intentionality detector (ID), the eye-direction detector (EDD), the shared attention mechanism (SAM), and the theory of mind mechanism (ToMM) (Figure 1). The author claims the modularity to be a necessary part of the model, as different clinical diagnoses can be explained by deficits in specific modules. The ID interprets self-propelled motion of entities in terms of its desires and goals. The EDD specializes in detecting eyes or eye-like stimuli, recognizes the direction of the gaze, and enables the mental attribution of the ability *to see* an observed entity. The purpose of the SAM module is to integrate the two types of information provided by the ID and EDD. This module already allows humans to determine whether another entity has the same target of visual attention. The ToMM module builds on the SAM module and achieves two goals: First, it allows inferring mental states in others from their observable behavior. Second, it allows us to generate explanations for observable behavior by integrating these hidden mental states into theories [Langton et al. 2000]. ID and EDD form dyadic representations (e.g., a cat chases a mouse (ID), or a cat sees a mouse (EDD)). The SAM module, however, builds triadic representations that are not possible only in the ID and EDD (e.g., I see a cat that chases a mouse). Finally, the ToMM module is able to represent the full range of mental state concepts. These are referred to as *M-Representations* and enable descriptions of mental states, where an agent has an attitude toward a proposition (e.g., Johnny believes that “the money is in the biscuit tin.”). There is research that builds on this model in the fields of clinical, developmental, and comparative psychology (where the latter studies the mental processes of non-human animals).

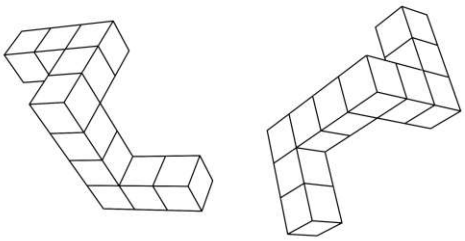


Figure 2 An example of a mental rotation task, adapted from [Just and Carpenter 1976].

2.2 Procedural Model of Joint Attention

Another approach to explain joint attention is to categorize processes involved in a successful joint attention event. From observations in infants, the two core processes are *responding to joint attention* (RJA) and *initiating joint attention* (IJA) [Mundy and Newell 2007]. RJA refers to the ability to follow the direction of the gaze and gestures of others. This allows to establish a common point of reference. IJA describes an infant's ability to use gestures and eye-contact to direct the attention of others. Targets of attention are either objects, events, or the infant themself. Clinical research shows that developmental deficits arise in either of these two processes separately. Comparative studies in non-human animals show that animals have the capacity for one of these processes, while little to none for the other. Chimpanzees, for example, can respond to, but rarely initiate joint attention [Tomasello et al. 2005].

2.3 Eye-Mind Hypothesis

The gaze occurs first to gather information, while it also signals information to observers, either intentionally or unintentionally. Just and Carpenter [1976] introduced a simple, yet powerful idea, namely the "eye-mind hypothesis." At that point in history, cognitive psychologists strived to understand what was then called the *central processor* of the human mind. Their experiments involved eye-tracking while performing mental rotation of Tetris-block-shaped three-dimensional objects (Figure 2) as well as checking whether displayed sentences correctly described the content of pictures next to them. The authors discovered relations between the ongoing mental operation and the gaze fixation target.

In summary, they found empirical evidence that the "locus of eye fixations reflects what is being internally processed" and that the "locus of the eye fixation can indicate what symbol is currently being processed" [Just and Carpenter 1976,

p. 53]. The term *symbol* indicates a mental content or entity, something one can think about. For example, when thinking about your favorite mug, your mental representation of that mug is a symbol.

However, there are limits to the eye-mind hypothesis: Webb and Renshaw [2008] argue that the eye-mind hypothesis is more likely to hold, when a person is performing a visual task, as opposed to pure cognitive tasks or tasks involving modalities other than the gaze.

2.4 Types of Gaze Behavior

As discussed in previous sections, there is strong evidence of some connection of the mental focus of attention and the current gaze target. In situations where a potential interaction partner is present, there are several plausible gaze targets. Looking at objects or specific locations other than the interaction partner is referred to as the *deictic gaze*. When two interaction partners are attending to each other's gaze it is called *mutual gaze*, colloquially eye contact. *Gaze following* is the action of attending to the gaze of the interaction partner, detecting their gaze direction, and then focusing their own gaze onto the stimulus that is being attended by the partner. Kaplan and Hafner [2006] also disambiguated the state of joint attention from gaze events that appear similar, but have a lower degree of coordination: 1) Simultaneous looking at an object that is triggered by a "pop-out" effect or salient event; 2) Coincidental simultaneous looking at the same object; 3) Gaze following of one agent, while the other pays no attention to the fact that they are being observed; 4) Coordinated gaze at the same object, but attention to different aspects of it (e.g., action intent (like playing with it), or aspect (like color)).

Gaze also plays a large role in pure conversation settings. For example, staring at the other person is often uncomfortable, unnatural, and does not lead to a smooth conversation experience for either participant. Therefore, gaze aversion is often equally important and serves different roles: First, it regulates the intimacy of a conversation. Secondly, it is utilized for turn-taking in a conversation. Gazing at the addressee after an utterance while being silent indicates that the other person should take the floor. Thirdly, averting the gaze indicates cognitive effort. Thus, a speaker can signal that they are not yet done with their turn, even though they are currently silently formulating a statement in their mind [Andrist et al. 2014].

3 Joint Attention in Human-Robot Interaction

An envisioned goal for Social Robotics is close collaboration between humans and robots, reaching beyond humans and robots working on different subtasks that lead to a common end result (e.g., pick-and-place robots in production). Actual collaboration between humans and robots is a sequence of shared actions toward a shared goal and requires coordination [Kolbeinsson et al. 2019]; in other words, joint attention as employed in the sofa moving example mentioned in the introduction. In our work, we explicitly focus on human-robot interaction (HRI) use cases surrounding object manipulation (e.g., picking up objects) and exclude settings with a stronger social focus.

There is no definitive theoretical model for joint attention on a robot. For implementation purposes, one approach is to view the desirable input-output relation for a given scenario as the requirement and use whichever technique is available and achieves the result. For example, a human and a robot can both generate plans for solving a given problem, but their specific methods can differ.

Additionally, Krämer et al. [2011] argued that the width and depth of human coordination capabilities in social contexts will be out of reach for technological systems in the foreseeable future (although constant progress is being made). We must instead direct our attention to artificial intelligence (AI) research and look for feasible components that solve simplified problems or help with a small part of the problem.

The authors split the problem of developing a ToM for Social Robotics into a micro (actual interaction), meso (relationship building), and macro level (roles and persona). On the micro level, they associate ToM, perspective taking, shared intentionality, and common ground. Common ground refers to mental content of which all interaction partners know that this content is known by everyone.In relation to these levels, our work addresses a joint attention implementation on the micro level, excluding considerations on the meso and macro level.

3.1 Implementing Joint Attention for HRI Tasks

HRI research has produced several results regarding joint attention implementations on robots. These include the capability of drawing attention to another reference point, as well as establishing, monitoring, and ensuring joint attention during an interaction. The interaction settings are either conversational with different points of interest in the environment or physical such as object handovers or other object manipulations.

These scenarios differ from pure conversational settings between a human and a robot. Typically, joint attention HRI settings involve at least another object, location of interest, or event besides the two agents. The human and robot both measurably focus their attention on this third entity or even physically interact with it. Imai et al. [2003] proposed an HRI joint attention mechanism in 2003. They presented the difficulty of drawing a person's attention to another reference point. This includes how to make a person understand the communicative intention of the robot, and how to deal with the person's attention status. They implemented the pointing and gazing functionality on a humanoid robot, enabled the robot to perform the mutual gaze, and represented the person's attentional focus as a spatial coordinate. They conducted an experiment, where the robot acted as a presenter of a scientific poster to a human participant. Results indicate that humans looked more frequently at the poster, when the robot displayed the proposed attention mechanism.

Huang and Thomaz [2010, 2011] extended the Responding and Initiating Joint Attention (RJA, IJA, Chapter 2.2) model by an explicit Ensuring Joint Attention component (EJA). The EJA component in their framework encapsulates the ability to monitor another's attention to verify that joint attention is reached and maintained. They describe a cannonical joint attention episode between two agents comprising five steps: 1) Connection of two agents, where they become aware of one another and anticipate an interaction; 2) Joint attention request by the initiating agent, where it focuses the attention on a third object and uses communicative channels such as pointing, gesture, and voice; 3) Joint attention response, where the other agent also focuses on the third object; 4) Monitoring, where the initiating agent ensures joint attention by switching the focus between the other agent and the referential focus; 5) Joint attention is reached, the interaction continues. The authors equipped their social robotic platform with a finite state machine, a procedural representation of the described joint attention episode. The perception capabilities of the robot included face detection, marker detection to perceive pointing actions, and speech recognition for a few phrases, which were used to check the attentional state of the human interaction partner. The humanoid robot had a movable head with two degrees of freedom and eyes with two degrees of freedom, as well as movable arms for pointing and a speaker for verbal communication. The authors conducted several experiments. In the first one, the robot had to show that it can respond to joint attention, by attending to objects that the humans pointed at. In the other experiments, which were video-based, the robot had to direct the attention of a human to a presentation as a tour guide, ensure attention while delivering a verbal message and while giving directions. The overall result indicates that robots with their joint attention implementation yielded better results in the responding to pointing actions task, and were considered more nat-

ural in the video-based experiments. Huang and Thomaz [2010] mentioned, that it is unclear how to design the specific timings of the EJA component.

Pereira et al. [2019] created an autonomous gaze system for the Furhat robot (a mounted mannequin head with an animated video-projected face) for a puzzle-like spatial reasoning task conducted on a tabletop. Their attention system is split into a proactive and a responsive gaze layer with different priority levels. Gaze events of higher priority override those with lower priority. The timing of gaze shifts is uniformly sampled from predefined ranges. The human participant, task objects, and the surrounding environment (for gaze aversion) are possible gaze targets. The proactive layer handles the gaze related to the speech acts of the robot (eye contact, IJA at task objects) and idle gaze behavior through gaze aversion. In the responsive layer, user speech activity and a detected mutual gaze led to a mutual gaze, while gaze tracking and object tracking was used for RJA events to gaze at objects. The system was then used to engage with the user during the task, comment on their progress and provide hints for the correct move. In a user study, self-reported data suggested that the robot with both responsive and the proactive layers was perceived as more socially present than the robot with only the proactive component, as only the former was able to react to the user and thus engage in joint attention.

Joint attention capabilities have also been shown to improve collaborative physical tasks like handovers in HHI [Frankel et al. 2012], but also HRI. Grigore et al. [2013] created a two layer architecture for physical robot-to-human handover tasks for a humanoid robot. The first layer represents the physical state of the handover as a Hidden Markov Model with the states “Robot pick up,” “Robot hold,” “User grap,” and “Robot not hold.” These states, however, are only estimated by the current and torque values measured in the robot hand. A higher-level layer was then added that serves as an additional safety check to release a grasped cup to the human under the right conditions. The authors observed that human users performed a sequence of actions in a successful handover: browsing the environment, looking at the target cup, (optionally looking at the cup repeatedly), and finally grasping the cup. The second layer registers the gaze pattern of the human by monitoring the head direction. Only if the described gaze pattern is detected before registering a grasp attempt, the robot releases the cup. The extension of the handover architecture has been empirically shown to result in fewer unsuccessful grasp attempts.

Similarly, Moon et al. [2014] compared HRI handover scenarios with varied humanoid robot gaze behavior. In an HHI handover study they detected two gaze patterns of the agent handing over the object: The shared attention gaze is gaze-directed at the projected handover location. In addition to this behavior, a turn-taking gaze pattern occurs sometimes, which consists of establishing eye contact

while reaching out. These findings were implemented in a humanoid robot, which resulted in the experimental conditions of no gaze (baseline), shared attention gaze, and the shared attention gaze plus turn-taking cue. The authors found that human users reached for the handover object earlier in the two gaze conditions, and reported a trend of self-reported preference for the turn-taking behavior over the other two conditions.

3.2 Planning for Joint Human-Robot Interaction

As Baron-Cohen [1994] mentioned, humans are expert mind readers. Hence, when a human observes another human in an everyday situation, the observer most likely forms an idea about what the observed person is trying to achieve with their current actions. For example, if you see someone in a kitchen opening the cupboard drawer containing all the mugs, you will probably already think about which drink they want to consume, while all they did was simply opening a drawer. Notable, it is quite possible that the observed person will do something different, but our experience tells us that getting a drink is the most probable goal given such an observation. One research direction on Joint HRI is to explore methods for simulating this human capability, namely AI planning.

We distinguish between symbolic and subsymbolic planning: In a formal language, symbols are atomic tokens of a language. This means they cannot be split into smaller units of meaning. Symbols are manipulated with some kind of procedure to build more complex expressions. This is (mostly) comparable to our spoken language with its single tokens, such as “cat,” “in,” and “tree.” From these tokens one can build expressions “cat in tree” or “tree in cat.” One of these makes more sense from our experience than the other, but both are correct expressions in our language. In turn, the expression “cat tree in” would not be considered as part of our language. There is simply no valid symbol manipulation sequence that can generate this expression. Nevertheless, symbols alone do not have any meaning in themselves, and the problem of assigning symbols to references in the physical or social space is referred to as the *symbol grounding problem* [Harnad 1990; Coradeschi et al. 2013]. In contrast, subsymbolic planning involves a more direct representation of the problem. Consider a map where one must find the shortest route between two points. There are no tokens that are manipulated, just path finding reasoning with the data provided by the map.

Generally, subsymbolic planning is often used for collaborative problems such as social navigation (i.e., safely moving through a crowd of people [Mirsky et al. 2021]) or human-robot handovers, where the problem is represented and solved in a task space like the Euclidean space of a suitable dimension. For more ab-

stract or high-level planning problems, however, a symbolic approach makes the problem formulation more compact. In this book chapter, we focus on such representations.

Before formulating the problem itself, however, we must consider our underlying assumption, namely the rationality of all involved agents. Broadly, this means that an agent would rather perform an action that results in a benefit to them, rather than harm. In the frame of the problem definition, the question is how to define a cost function, or even how to know that optimizing the *expected* cost for a problem is even the right thing to do [LaValle 2006]. Assigning reward (or cost) values to certain outcomes of a decision process may be intuitive. These may be of a monetary value, or of a more subjective value, like choosing between washing the dishes or sweeping the floor. Thus, every action is assigned a reward value. If the action outcome is stochastic, then a reward distribution is assigned to each action. An example of this is a game where an agent chooses between receiving 1000 € or letting a coin flip decide whether they receive 2000 € or nothing. Although the expected value of both actions is the same, most people will have a preference for one or the other, depending on their inclination toward gambling. Thus, using the expected value alone is insufficient to model the preferences of agents. This is solved by deriving a so-called utility function for all action outcome distributions. For a utility function to exist, a rational agent must be able to provide a consistent ranking of different probability distributions over outcomes according to the axioms of rationality [LaValle 2006]. Thus, each action outcome is assigned an utility value. Finally, a cost function can be derived from the utility function.

Markov Decision Processes (MDP) can be used to solve problems in sequential decision theory [LaValle 2006], where agents repeatedly chose actions according to their current state. A single agent MDP is defined by 1) a non-empty ***state space X***, which is a finite or countably infinite set of states; 2) for each $x \in X$ a ***finite, non-empty action space U(x)*** with a ***termination action*** (it is applied when reaching a goal state); 3) a finite, non-empty ***nature action space*** $\Theta(x, u)$ for each $x \in X$ and $u \in U(x)$ (a ***nature*** decision maker represents uncertainty in the action outcome); 4) a state transition function f that produces a state, $f(x, u, \theta)$, for every $x \in X$, $u \in U$, and $\theta \in \Theta(x, u)$; 5) a set of ***stages***, which is either infinite or set to a fixed, maximum stage (i.e., how many sequential actions can be taken before the problem must be solved); 6) an initial state $x_I \in X$; 7) a goal set $X_G \subset X$, and 8) a stage-additive cost functional L. The goal of the agent is to find a plan to reach a goal state from the initial state. Because there are stochastic state transitions, a policy $\pi : X \rightarrow U$ must be found for all $x \in X$ that minimizes the cost. Alternatively, π can be a mapping from a state to a probability distribution over the action space. Then, this corresponds to a *randomized* instead of a *deterministic* strategy.

Markov chains are a simplification of this model without an explicit decision maker. Nature determines the outcome of the next state alone. Markov chains are used to model stochastic processes and, like MDPs, fulfill the Markov assumption (equation 1). $X_1, X_2, \ldots, X_t$ denotes the sequence of random variables up to timestep t, where the outcomes are $x_i \in X$. This means that only local information, and not the entire history of the process is used to determine the probability of the next state transition.

$$Pr(X_{t+1}) = x_{t+1} | X_1 = x_1, X_2 = x_2, \ldots, X_t = x_t) = Pr(X_{t+1}) = x_{t+1} | X_t = x_t) \tag{1}$$

Generally, artificial agents have some sensing capability to determine the current state they are in. However, due to nature, sensor errors can occur. This leads to anoher type of uncertainty, besides stochastic state transitions, namely state uncertainty. This means that the agent does not know for sure whether it is in a single current state $x_t \in X$, but holds a ***belief*** about the current state, expressed as a probability over X. Including this belief into planning lifts the problem formulations from the state space into the state belief space.[2]

For joint action scenarios, it is important to model more than one active decision maker. This leads to the inclusion of the game-theoretic concept of the *two-player nonzero-sum game* [LaValle 2006]. One formulation is to extend the MDP definition by another agent. Herein the two agents (players) P_1 and P_2 have their respective action spaces U_1 and U_2. In zero-sum games, there is only one cost function $L : U \times V \to \mathbb{R} \cup \infty$, which one player regards as reward, and the other player as cost. In the nonzero-sum game, however, each player has a different cost function (like L), namely L_1 and L_2. Both players now aim to minimize their costs according to their respective cost function. Thus, in such games different degrees of cooperation can be formulated, from total cooperation to a zero-sum game. This formulation can be lifted to sequential games on game states by expanding the MDP definition by another player.

In symbolic planning problems, if the planning problem uses deterministic action outcomes, a wide-spread approach in robotics is to employ *classical planning*. A *classical planning domain* (i.e., a *state-transition system*) is a triple $\Sigma = (S, A, \gamma)$ or a 4-tuple $\Sigma = (S, A, \gamma, cost)$. S is a finite set of possible ***states*** of a system. A is a finite set of ***actions*** that an actor can perform. $\gamma : S \times A \to S$ is a partial function called the ***state-transition function***. When $\gamma(s, a), s \in S, a \in A$, is defined, then a is ***applicable*** in s, and $\gamma(s, a) \in S$ is the outcome of the action. $cost : S \times A \to [0, \infty)$ is a partial function with the same domain as γ, defining a metric, which is to be

2 Literature presented in this chapter as well as our contribution only concerns planning in state space.

minimized, such as the monetary cost or time. In this kind of representation, there are the assumptions of a *finite, static environment*, *no explicit time* (except the cost, if it is to be interpreted in this way), and *no concurrency*, indicating that actions cannot be performed in parallel. Actions are *deterministic*, which means that the outcome of an action is known with certainty [Ghallab et al. 2016].

In the formulation above, there is a finite set of states ($S = (s_0, s_1, \dots)$) with no specific relation to one another. A more succinct way of describing states is by using *state-variables* (*predicates*) and *objects*. Hereby, states are defined as specific instantiations of these state-variables. These state-variables can use objects as arguments. A concrete example is the planning domain `blocksworld` in the *Planning Domain Definition Language* [Fox and Long 2003] (PDDL), which is a formal planning language that is commonly used for robotic tasks that involve planning in semantic domains. It is an approach to encode a classical planning problem, derived from previous formal languages like the *Stanford Research Institute Problem Solver* (STRIPS) [Lifschitz 1987]. A PDDL problem is encoded by a domain and a problem instance, where the domain describes the state-variables and operators, which are uninstantiated action templates. Once an operator is given parameters, it is called an action. Operators, like `pickup`, are defined with objects as possible parameters (`?ob`), preconditions, and effects. Only when the preconditions are met in the current state, the action is performed by applying the effects of the action on it. This is done by adding and/or removing predicates from a state. The problem instance describes the existing objects, the initial state, and the goal. The solution represents a plan, which solves the problem. There are PDDL versions that allow durative and concurrent actions, continuous and conditional effects, etc., however, we disregard these options for simplicity.

3.3 Plan Recognition in Classical Planning

Classical, symbolic AI planning is an approach to endow a robot with a planning capability suitable for joint HRI situations. However, it is only a part of the solution. A robot must also be able to infer the goal and plan of the interaction partner. To this end, classical planning plan recognition is employed [Ramírez and Geffner 2009, 2010; Sohrabi et al. 2016]. An advantage of this approach is the reuse of the planner that the robot uses to generate its own plans. The plan recognition problem is formulated as a triple $T = \langle P, G, O \rangle$, where P is a planning domain, G is a set of goals, and O is a sequence of observed actions. When the sequence O ends in a state that is a goal, the goal recognition is trivial; however, when the observation ends in a state that is not a goal, the problem is to predict which is the most likely goal, to rank these goals with regard to their relative probabilities, or to assign probabilities to the different goals. Various approaches have different

ways of executing this, but their commonality is to transform the original planning domain to accommodate the observations and subsequently compare the cost of different plans. Different plans are generated for a single goal, e.g., one that satisfies the observations and one that does not. When the cost of adhering to the observation for a goal is significantly higher than reaching the goal without doing so, that goal is probably not likely to be the actual goal of the observed actor. This builds on the assumption of *rationality* of an agent, i.e., that one attempts to fulfill their desires in an effective and efficient way.

3.4 A Benchmark in HRI for Joint Action

Situations that are simple and intuitive to solve for a human team, such as building a specific tower out of wooden blocks on a table, prove to be complex and difficult for current joint attention research. Therefore, this setting - a human and a humanoid robot who attempt to build a block tower - is used as a recurring scenario in joint action research [Johnson et al. 2009; Schulz et al. 2018; Barchard et al. 2020; Jensen 2021].

Pure plan recognition research often only treats problems that are already formulated in formalisms like PDDL. Similarly, the problem formulation of plan recognition does not deal with the continuous coordination effort that is necessary in joint attention situations. Devin et al. [2017] combined classical planning in the block world domain with the demands of joint action problems. In their study, they set up a joint action scenario with a human participant and a PR2 robot[3] (Figure 5, left). The PR2 robotic platform was equipped with several optical sensors and two arms with pincer grippers. The setup includes fiducial markers on the blocks to facilitate their recognition. The robot was able to perceive the world state (i.e., the current arrangement of blocks) and manipulate the blocks.

The robot and the human participant have a shared goal. They stand on opposite sides of a table and attempt to build a specific block tower with blocks lying on the surface. Howeer, each agent is only able to reach some of the blocks, hence they must collaborate. To introduce another challenge, there is not one single fixed sequence that results in the correct block tower (Figure 3). For example, there are two places for putting the red blocks and each actor has access to one of the two red blocks. They need to coordinate who picks which placement spot. The following difficulty arises when the agents must place the block stack green-blue-green. Again, each actor has access to only one green block. Thus, the actors must coordinate who places the first green block.

3 https://robots.ieee.org/robots/pr2/, Image source: https://www.wevolver.com/wevolver.staff/pr2

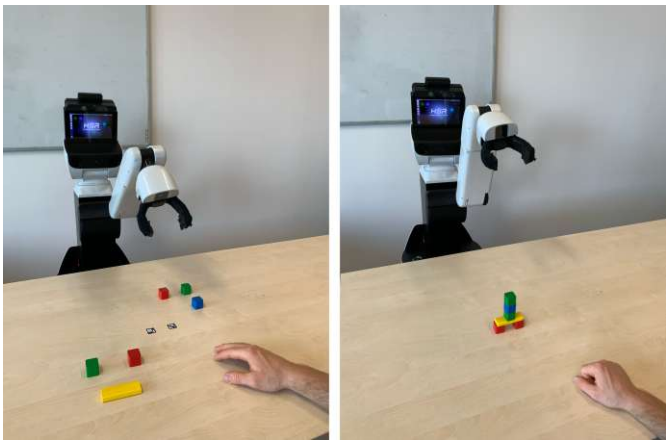


Figure 3 Joint action task described in [Devin et al. 2017]. Left: Initial configuration. Right: Goal State.

The authors approach this scenario as a multi-agent planning problem. The robot finds plans by modeling three discrete actors (itself, the human, and a fictitious *X agent*) who can place the blocks. In valid plans, actions that are assigned to the *X agent* mean that either of the two actors human or robot will perform the action. Notably, in the example above, there could be multiple open actions at once, e.g., placing the two initial red blocks in the center. In the shared plan, when the next necessary step is an action performed by the human, the robot waits for its completion. When the next necessary step is a robot action, the robot performs it. However, whenever an action is assigned to the *X agent*, the robot has different approaches for enacting this shared plan, namely acting lazily (i.e., waiting for a specified amount of time and watching whether the human will perform the action) or in a hurried way (i.e., the robot always attempts to immediately perform an *X action*. Furthermore, agent assignments can change during the plan execution, such that the plan must be recalculated after each step. For example, when one actor places the first green cube, the placement of the second cube is no longer an *X agent* action, as only the other agent has a green block left. This demonstrates the complexity of this simple collaborative block world problem as it already exposes numerous interesting and difficult aspects of joint action and requires further research effort. Thus, to establish a standardized scenario, Clodic et al. [2017] propose a joint action scenario similar to Devin et al. [2017]. Their goal was to facilitate finding answers to the following questions: “What knowledge does a robot need to have about the human it interacts with [...]?”; “What information should the human possess to understand what the robot is doing and how the robot should make this information available [...]?” [Clodic et al. 2017, p. 2] The proposed simple HRI scenario has the following setup and assumptions:

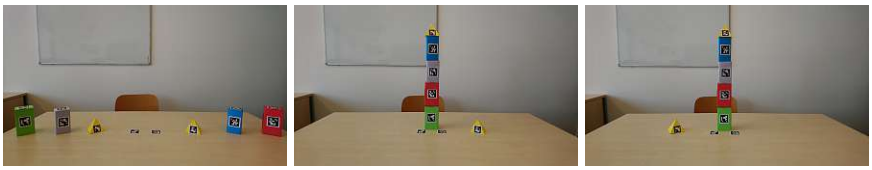


Figure 4 Left: Initial configuration. Middle and Right: The two possible goal states.

The common goal of the human and robot is to build a stack of four blocks in a specified order with a pyramid on top. They are on opposite sides of the table and face each other. Each agent has access to two of the four blocks. There are two pyramid pieces, one on either agent's side of the table. Only one of the two agents is supposed to place the pyramid piece at the end of the action sequence. The agents are restricted to the actions of the block world domain, plus a handover action, and a possibly support tower action.

Figure 4 illustrates the initial and the possible goal states. Both agents are assumed to perceive the current world state and thus are able to locate objects and assess their reachability by either agent. Finally, each agent is able to observe actions of the other.

4 Toward a Gaze Mechanism for Joint Actions

As described above, one of the two core questions posed by Clodic et al. [2017] is *how a robot should signal information that is important to the human in order to enable smooth collaboration*. We argue that the gaze is a useful modality for this specific benchmark task even for robots, as it is highly intuitive for humans to interpret, and is perceived constantly without being bothersome (in contrast to continuously verbalizing information, for example). It is furthermore potentially easier to perform than other non-verbal behavior, e.g., pointing.

Conveniently, common mobile service robotic platforms such as the PR2 by WillowGarage or the Toyota Human Support Robot[4] (HSR) (Figure 5) have head-like extensions with two degrees of freedom that house forward-facing optical sensors. Therefore, the head orientation represents in fact the direction of gaze. Social humanoid robotic platforms, such as Pepper from Softbank Robotics[5] or Nao[6] (Figure 6) have the same degrees of freedom in their heads and have al-

4 https://robots.ieee.org/robots/hsr/, Image source: https://developer.nvidia.com/embedded/community/reference-platforms/toyota-hsr

5 https://www.softbankrobotics.com/emea/en/pepper

Figure 5 Two domestic service robots. Left: Toyota Human Support Robot (HSR). Right: PR2 by WillowGarage.

ready been used in gaze related HRI studies. Research has shown that their head orientation communicates attention [Breazeal et al. 2005; Takayama et al. 2011] and is interpreted as gaze by human participants. We, therefore, propose that the gaze in the joint action benchmark will significantly smooth the interaction between the human and the robot, as it has previously in the different communicative HRI settings surveyed by Admoni and Scassellati [2017].

4.1 Comparison of Human-derived Gaze Mechanisms

It is important to model the gaze behavior of domestic service robots in a way that it primarily does not impede their functionality, and secondly serves a communicative purpose in joint attention and joint action situations. The human gaze is very effective at doing both simultaneously. During object manipulation tasks, humans gaze at task-relevant objects and locations [Hayhoe and Ballard 2005; Pelz et al. 2001]. This behavior is a rich source of information for an interaction partner in collaborative scenarios. In the ideal case, a robot would use its gaze to improve its belief about the current world state, as well as utilize the communicative aspect of gaze. Therefore, a model of the human gaze in joint action tasks can be used as an initial heuristic. The most important characteristics of such a model are the gaze locations and timings, i.e., when to look at what. Another, perhaps less important factor, are the transition dynamics, i.e., which animation profile is exhibited by gaze transitions.

When implementing a gaze model for a robot that interacts with another actor and objects in its environment during a joint task, the question of *when the robot looks at a specific gaze target needs to be addressed*. More specifically, which se-

6 https://www.softbankrobotics.com/emea/en/nao

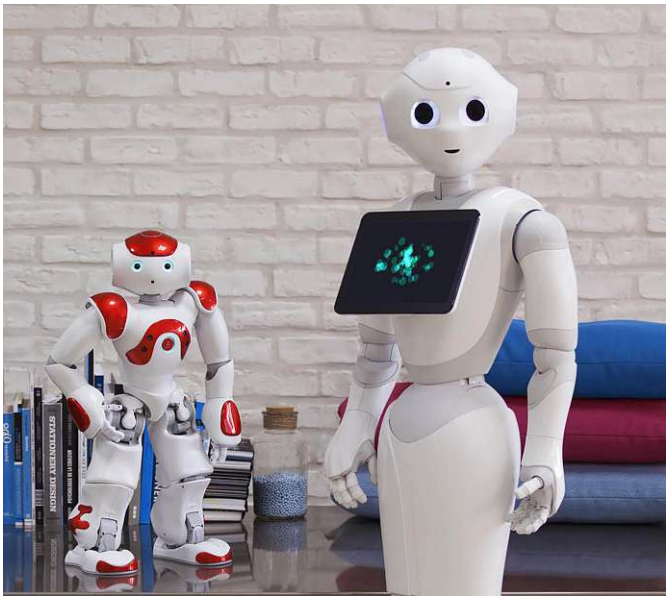


Figure 6 Two social humanoid robots by Softbank Robotics. Left: Nao. Right: Pepper.

quence of gaze targets and fixation durations communicates the attentional (gaze) focus of the robot to the human actor? We assume that the gaze is divided between the objects the robot manipulates itself, the object manipulations of the human partner, and the human's hands and face. The gaze at the objects that the robot wants to manipulate is (at least at some point in the process) necessary for the proper execution of the planned action. Thereby, the robot communicates its own attentional focus through gaze. The gaze at the object manipulations by the human is necessary to assess the current world state. The gaze at the face of the human is necessary to ensure the joint attention status. Similarly, at each point, the gaze of the robot could be interpreted by the human to draw conclusions about the attentional state of the robot.

This might seem to overly complicate the block stacking benchmark task, however, it represents only an initial step to solve more difficult scenarios. Examples of these include tasks with more than two actors, and tasks that include more movement, such that not each important location of attention is captured in a single camera angle, for example when objects are positioned further apart, when actors do not face each other all the time, or when objects are occluded.

4.2 Modeling the Sequence of Gaze Targets

Next, we discuss how to create a gaze model for the above-mentioned tasks. Lehmann et al. [2017]; Acarturk et al. [2021] employed a specific methodology for creating a gaze controller specifically for gaze aversion in conversational settings. They recorded two eye-tracking datasets in dialogs between two humans, where one participant was the interviewer and the other the interviewee. One dataset was generated from the view of the interviewer, the other one from the view of the interviewee, using a wearable Tobii Glasses 2[7] eye-tracker. For each interview perspective they used a sequential data mining method to derive the most common gaze shifts, where the following gaze targets were encoded: the face of the dialog partner (referred to as *gaze contact fixation* by the authors), and gaze aversion directions relative to the position of the face (down, up, left, right, and diagonal directions).

More importantly for this book chapter, stochastic models are also used to model gaze sequences. (First order) Discrete-Time Markov Chains (DTMC) describe sequences of gaze directions using the Markov property assumption (Equation 1, Section 3.2), i.e., only the previous gaze target determines the probability of the next gaze direction and the possible states are in the set
$\Omega = \{center, up, down, left, right, up{-}left, up{-}right, down{-}left, down{-}right\}$.

A simplifying assumption was made, namely time-invariance, meaning that the probabilities do not change depending on the position in the sequence. This allows the gaze model to be represented as a Markov chain transition matrix of size $|\Omega| \times |\Omega|$. A cell matrix cell value p_{ij} represents the probability of changing the gaze from target x_i to x_j and the rows must sum up to 1.

The authors argued that a gaze controller producing such stochastic behavior will be helpful in HRI conversational settings. Further, they have future plans to validate this idea by implementing it on a humanoid robot and conducting HRI validation studies following the methodology of Andrist et al. [2014], where the proposed model with proper gaze timings was tested against a baseline with static gaze and a baseline with inverted timings (“anti-timings”). The study argued that both baselines should lead to a worse evaluation of the robot by the human interview partners than the proposed model.

This kind of gaze control is aimed at conversational HRI settings and has numerous useful applications, such as tour and info guidance, receptionist duties, etc. Mobile service robots such as the Toyota HSR can additionally perform object manipulation tasks and require gaze control for them, as argued above. Provid-

7 https://www.tobiipro.com/product-listing/tobii-pro-glasses-2/

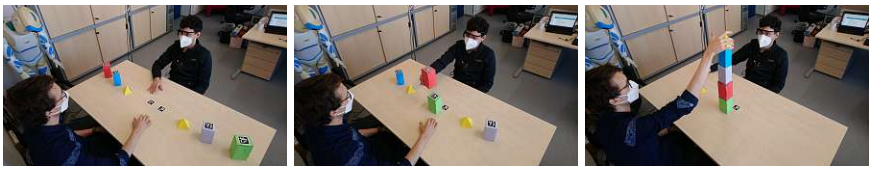


Figure 7 Gaze data capturing during the pilot study. Left: Initial position. Middle: Eye-tracked participant places a block from the reachable area. Right: Placement of the pyramid block. Both participants can place their pyramid, and after a negotiation phase, the other participant places the final piece.

ing a gaze controller for the joint action benchmark task described earlier is thus helpful to handle more realistic scenarios in the future.

5 Data Collection for our Stochastic Gaze Control

We describe how to adapt the procedures from Lehmann et al. [2017]; Acarturk et al. [2021] to a collaborative object manipulation task. In a pilot study, we recreated the block stacking task with the pyramid top presented in Clodic et al. [2017] (Figure 7). Two human participants sit opposite each other at a table. One of the two participants per trial wore a PupilLabs Core[8] [Kassner et al. 2014] eye-tracker with monocular eye-tracking.

We tested two pairs of participants ($n = 4$). Each pair conducted two trials. After the first trial, they swapped positions, such that each participant wore the eye-tracker in one trial. All participants were briefed by the experimenter. The participants were asked to read and sign an informed consent form. They were instructed to collaboratively build a specified tower (from bottom to top: green - red - lavender - blue - pyramid). Figure 4 depicts the view of the person wearing the eye-tracker. This person was instructed to act as if only the red block, blue block, and right pyramid is reachable for them. The person sitting opposite was instructed to act as if they can only reach the green block, the lavender block, and the left pyramid.

The participants were instructed to follow a set of rules: (1) Use only your right hand. The task was simple enough for humans, such that non-disabled persons can use their right hand even if it is not their dominant hand. (2) The right hand is supposed to always be above the table. (3) The left hand is supposed to be out of sight underneath the table. (4) Participants were asked not to rotate the blocks while moving them.

8 https://pupil-labs.com/products/core/

The participants were informed that this is not a test and that speedy execution is not important. Starting a grasping action while the other person is still placing their block was not forbidden. The blocks display fiducial markers facing the person wearing the eye-tracker and participants were asked to grasp the block in a way that does not occlude the markers. The placement position of the bottom block was also marked on the table with fiducial markers. These rules and restrictions were implemented such that the resulting behavior is similar to the one of a robot during such a task.

The two participants were asked to memorize and recite the correct block stacking sequence before the experiment to avoid execution mistakes and to limit gaze and other behavior that is not associated with shared plan execution. The participants were not allowed to discuss any strategy before the task and were not allowed to speak during its execution.

The participant wearing the eye-tracker is referred to as the *robot* (R), because the recorded gaze behavior is meant to be implemented on a service robot in the future. The other participant is referred to as *human* (H). X denotes the *X Agent* (X). The resulting interactions included only actions that were in accordance with the optimal plan:

```
(pickup H green) (place H green table) (pickup R red)
(stack R red green) (pickup H lavender) (stack H lavender red)
(pickup R blue) (stack R blue lavender) (pickup X yellow)
(stack X yellow blue)
```

Gaze behavior that results from these interactions thus depicts gaze behavior for smooth interaction without errors. During the last step, where the two agents need to negotiate who picks up their pyramid piece, gaze behavior indicative of negotiation will take place. The generalization is naturally only possible for an appropriately large sample size and only for populations with the same demographic properties. In this chapter, only a preliminary feasibility check with a small sample size is presented, and the obtained results serve as an exemplary outcome.

The goal of this experimental setup is to elicit successful collaboration and the corresponding gaze behavior in the person wearing the eye-tracker. Large-scale plan re-negotiations during the task must be avoided. Small-scale negotiations (i.e., resolution of *X agent* actions) fall within the capabilities of the planning formalism. This choice is motivated by the consideration of the full robot architecture: In problms that are more general than the chosen experimental setting, large-scale plan deviations might occur. However, after each action (planned or unforeseen), the visual sensors of the robot will detect the resulting world state, which will be used as the initial state to the planning problem. Then, a new shared plan will be calculated. This might result in a new planned sequence of actions. The robot

gaze controller always acts with respect to a determined plan, as described below in further detail. Thus, if a new plan is calculated, the gaze is adjusted according to the newfound plan. Plan changes occur due to unforeseen actions; however, this does not result in unspecified gaze behavior. The robot gaze always corresponds to the belief of the robot and visualizing the belief of the robot through gaze is the goal of this gaze controller.

During the trials, the strategy to overcome the ambiguity of who places the pyramid was always solved with the “turn-taking” strategy, where the person who placed the topmost rectangular block waits for the other person to place the pyramid. In our small sample, the placement of the pyramid occured either immediately or after a short period of inactivity.

For each gaze data sample, we conducted the following evaluation: Using fiducial markers[9], as well as (the partner’s) hand and face tracking [Lugaresi et al. 2019] allowed the recognition of these objects in the eye-tracked video. By defining a 100 pixel radius around each target, we distinguish eye fixations of the other person’s hand and face, as well as the placement location of the bottom block on the table, as well as all other blocks and pyramids. Furthermore, we encode fixations gazing at none of the above.

For each sample, a sequence of fixations is extracted from the gaze data, and we create a DTMC transition model by counting the transitions. In this scenario, this yields a 8×8 matrix (pyramids are counted as one object). The gaze targets are the face of the partner, the hand of the partner, the placement location on the table, the four blocks, and the two pyramids, which are counted as one object due to their interchangeability.

For this gaze controller, we disregard fixations that do not fall in the radius of any target. If a fixation falls on a spot in the visual field that is currently in the radius of more than one target, we count split transitions and mark more than one object as currently active, until the gaze falls on a single object again.

The aggregated model in Table 1 was derived with the gaze model for every sample. There are two possibilities of arriving at the probability values, which sum up to 1 per row: Either the frequency counts of the transitions are averaged per sample, and then the averaged matrices are added and again normalized per row. This is the variant we chose, since it leads to equal representation of each sample. Another method is to add all frequency count tables and only then normalize over the rows.

The controller can then be applied to create gaze behavior by choosing a basic timestep unit, e.g., one second (This varies with the task, and the robot embod-

9 https://april.eecs.umich.edu/software/apriltag

	Next Target						
Target	Face	Hand	Table	Green	Red	Lavender	Blue	Yellow
Face	0.12	0.12	0.29	0.17		0.17		0.13
Hand	0.13	0.23	0.02	0.22	0.11	0.11	0.07	0.11
Table	0.11	0.37	0.08	0.25	0.04	0.04		0.11
Green		0.30	0.05	0.24	0.14	0.05	0.17	0.05
Red	0.10	0.10	0.25	0.12	0.23	0.10	0.10	
Lavender	0.38	0.07			0.07	0.11	0.26	0.11
Blue		0.19	0.04	0.11	0.04	0.14	0.48	
Yellow	0.67	0.17	0.08	0.08				

Table 1 DTMC transition probabilities of eye-tracked locations.

iment.) and creating a gaze sequence by starting in a random or predetermined (e.g., face) state. The next state is always sampled with the probability weights of the row of the current state.

Further work is planned to split the gaze controller into two parts and to analyse whether the gaze behavior in the action phase (placement up to the last block) differs from in the negotiation phase (placement of either pyramid).

5.1 Creating a Gaze Controller for Time-Variant Scenarios

Table 1 indicates the specific objects the participants gazed at during the whole task duration. This neglects an important factor, namely the dynamic nature of the time-variant task. During the task, the world state is defined by the block arrangement and whether an actor is currently grasping a block. It is clear to both actors which block to grasp next (or whether to negotiate who should place the pyramid top). For the plan execution, the following block to be placed has another role to the actors of the current action than a block that has already been placed. Therefore, we annotate the video samples with the current state of the world, i.e., which blocks have already been stacked (neglecting whether a block is grasped or not). Thereby, we partition the set of blocks, pyramids and table placement location into sets of *past*, *previous*, *current*, *next*, and *future*. The *current* block is the one that must be picked up and placed at a specific point in time. The *previous* block is the block that was placed right before the current block. Prior to placing the first block, *previous* indicates the table placement location. The *next* block indicates the block to be placed after the current block. *Past* and *future*

	Next Target						
Target	Face	Hand	Past	Prev.	Curr.	Next	Future
Face	0.08	0.08	0.19		0.11	0.19	0.33
Hand	0.16	0.19		0.09	0.27	0.20	0.09
Past	0.11	0.11				0.78	
Previous			0.12	0.12	0.12	0.12	0.50
Current	0.20	0.35			0.19	0.22	0.03
Next	0.23	0.12	0.11	0.06	0.31	0.15	0.03
Future		0.33			0.50	0.17	

Table 2 DTMC transition probabilities of eye-tracked locations in their dynamic context of the plan execution.

blocks group blocks that have been placed before *previous*, and must be placed after *next*, respectively. The controller in Table 2 is derived with this dynamic assignment of object roles. Hence, we preserve the time-invariance assumption of the gaze controller with this transformation from block identities to temporal roles.

5.2 Future Work

We tested the described pipeline to derive a gaze controller with transition probabilities based on a larger sample size. Careful attention to the validity of the result must be paid, as numerous design choices have been taken in the aggregation method of the different study participants and filtering of fixations in single samples. Therefore, we propose a validation study, where a pre-programmed humanoid robot and a human participant perform the described task. The robot functions according to the same assumptions as the one described by Clodic et al. [2017]. The robot acts in two different conditions: It can place the final piece proactively (try to do it itself) or “lazily” (wait until the human places it). During the task, the robot exhibits gaze behavior in accordance with the gaze controller derived from the empirical data collection. There will be two baseline conditions, namely one where the robot does not display any gaze behavior at all, and another one, where the robot acts according to “anti-timings,” as in the study of Andrist et al. [2014].

For the gaze controller, there are numerous possible elaborations. For example, the state space of the temporal roles could be expanded by the belief of who

the believed actor of that action is. The state space would then be $\{past, previous, current, next, future\} \times \{robot, human, Xagent\}$. The robot gaze could thus vary when the robot believes that the human is about to perform the next action in contrast to when the robot believes that it is to perform the next action itself.

While the approach in Lehmann et al. [2017]; Acarturk et al. [2021], and Andrist et al. [2014] has worked in conversation settings, it is unclear how gaze processes with dynamic gaze targets are handled by a robot. As human-like object manipulation capabilities are the current goal of service robotics research, human-like gaze behavior in object manipulation tasks is also beneficial, as humans are known to actively seek out information that helps solve the current task. This approach has a counterpart in robotic vision, called *active vision* [Aloimonos et al. 1988]. Future research can make use of the derived gaze timings to more reliably focus on important aspects of a scene, according to the ongoing task.

6 Conclusion

In this chapter, we mainly focused on research in psychology and HRI on joint attention, although there are numerous other related interesting subfields that influence how to think about joint attention in service robotics.

In psychology, attention is studied in numerous different scenarios, such as sustained attention, vigilance, and other low-level models of attention. In developmental psychology, research on the autism spectrum disorder in infants and developmental robotics explore how social collaboration abilities develop and emerge in complex behavior from more simple prerequisites. Studies in neuroscience and psychophysics focus on the neurological processes leading to the attention phenomenon. Differential psychology studies how personality traits lead to different modes of attending to stimuli.

Similarly, for AI/robotics, there are numerous fields that deserve a mention in attention research. Visual attention is an inductive bias, often used in visual pattern recognition and machine learning research. Multi-agent reinforcement learning deals with the emergence of communication protocols between untrained agents and how they attend to each other to solve complex collaborative tasks. In different computational cognitive architectures, joint attention may be a feature that emerges from the dynamic interplay of different architecture components. In machine vision, object detection plays a critical role regarding which objects can be paid attention to. Only if an object is detected, segmented, or classified, it will be able to enter the center of attention. In planning and scheduling, there are

numerous different paradigms with many different frameworks, of which a single one was chosen as the focus in this chapter.

To summarize this chapter, first, structural and procedural models of joint attention from the psychological perspective were discussed. The special relation between ToM and joint attention was of particular interest. We then focused on gaze as the main sensory modality. Information gathered through gaze not only provides necessary information to calculate mental representations of one's surroundings, but it is also driven top-down to focus on areas that are crucial to form a coherent explanation. This gaze behavior can be a source of information for observers.

Second, we reviewed how these insights are used to create robotic implementations for different joint attention or joint action scenarios. The scenarios included conversations with locations of interest other than conversation partners or collaborative physical tasks with different manipulable objects.

Third, decision-theoretic and classical planning were reviewed for their use in such collaborative physical tasks. Special attention was paid to plan recognition and the usefulness of a benchmark (building a tower out of blocks) for joint action in HRI.

Finally, we proposed a method for learning a stochastic gaze controller for such tasks from data. The joint action benchmark of jointly building a tower was used as experimental foundation. We presented a method to preserve the time-invariance assumption of the stochastic controller by assigning temporal roles to objects. These roles are assigned dynamically by checking the current world state and the shared plan. This was followed by an outlook on future research needed for the development of a novel gaze mechanism for joint actions in HRI.

Clearly, the work presented in this chapter only is a building block to a significantly larger research problem, namely how to enable humans and robots to succeed in dynamic collaborative tasks. However, it also demonstrates that attention is a topic that must not only be considered relevant for HRI research, but for the entire robotics field.

Bibliography

Cengiz Acarturk, Bipin Indurkya, Piotr Nawrocki, Bartlomiej Sniezynski, Mateusz Jarosz, and Kerem Alp Usal. 2021. Gaze aversion in conversational settings: An investigation based on mock job interview. *Journal of Eye Movement Research* 14, 1.

Henny Admoni and Brian Scassellati. 2017. Social eye gaze in human-robot interaction: a review. *Journal of Human-Robot Interaction* 6, 1, 25–63.

Nameera Akhtar and Morton Ann Gernsbacher. 2007. Joint attention and vocabulary development: A critical look. *Linguistics and Language Compass* 1, 3, 195–207.

John Aloimonos, Isaac Weiss, and Amit Bandyopadhyay. 1988. Active vision. *International Journal of Computer Vision* 1, 4, 333–356. https://doi.org/10.1007/BF00133571

Sean Andrist, Xiang Zhi Tan, Michael Gleicher, and Bilge Mutlu. 2014. Conversational gaze aversion for humanlike robots. In *2014 9th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 25–32.

Kimberly A Barchard, Leiszle Lapping-Carr, R Shane Westfall, Andrea Fink-Armold, Santosh Balajee Banisetty, and David Feil-Seifer. 2020. Measuring the perceived social intelligence of robots. *ACM Transactions on Human-Robot Interaction (THRI)* 9, 4, 1–29.

Simon Baron-Cohen. 1994. How to build a baby that can read minds: Cognitive mechanisms in mindreading. *Cahiers de Psychologie Cognitive/Current Psychology of Cognition,* 13(5), 513–552.

Simon Baron-Cohen. 1997. *Mindblindness: An essay on autism and theory of mind.* Cambridge: MIT Press.

Cynthia Breazeal, Cory D Kidd, Andrea Lockerd Thomaz, Guy Hoffman, and Matt Berlin. 2005. Effects of nonverbal communication on efficiency and robustness in human-robot teamwork. In *2005 IEEE/RSJ international conference on intelligent robots and systems.* IEEE, 708–713.

Aurélie Clodic, Elisabeth Pacherie, Rachid Alami, and Raja Chatila. 2017. Key elements for human-robot joint action. In *Sociality and Normativity for Robots. Studies in the Philosophy of Sociality*, Hakli R, Seibt J (eds). Springer, Cham. 159–177 https://doi.org/10.1007/978-3-319-53133-5_8

Silvia Coradeschi, Amy Loutfi, and Britta Wrede. 2013. A short review of symbol grounding in robotic and intelligent systems. *KI - Künstliche Intelligenz* 27, 2, 129–136. https://doi.org/10.1007/s13218-013-0247-2

Sandra Devin, Aurélie Clodic, and Rachid Alami. 2017. About decisions during human-robot shared plan achievement: Who should act and how? *International Conference on Social Robotics*, 453– 463.

Maria Fox and Derek Long. 2003. PDDL2. 1: An extension to PDDL for expressing temporal planning domains. *Journal of Artificial Intelligence Research* 20, 61–124.

Richard M Frankel, Mindy Flanagan, Patricia Ebright, Alicia Bergman, Colleen M O'Brien, Zamal Franks, Andrew Allen, Angela Harris, and Jason J Saleem. 2012. Context, culture and (non-verbal) communication affect handover quality. *BMJ Quality & Safety* 21, Suppl 1, 121–128. DOI: 10.1136/bmjqs-2012-001482

Malik Ghallab, Dana Nau, and Paolo Traverso. 2016. *Automated planning and acting.* Cambridge University Press. https://doi.org/10.1017/CBO9781139583923

Elena Corina Grigore, Kerstin Eder, Anthony G Pipe, Chris Melhuish, and Ute Leonards. 2013. Joint action understanding improves robot-to-human object handover. In *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems.* IEEE, 4622–4629.

Stevan Harnad. 1990. The symbol grounding problem. *Physica D: Nonlinear Phenomena* 42, 1-3, 335–346.

Mary Hayhoe and Dana Ballard. 2005. Eye movements in natural behavior. *Trends in Cognitive Sciences* 9, 4, 188–194.

Chien-Ming Huang and Andrea L Thomaz. 2010. Joint attention in human-robot interaction. In *2010 AAAI Fall Symposium Series.*

Chien-Ming Huang and Andrea L Thomaz. 2011. Effects of responding to, initiating and ensuring joint attention in human-robot interaction. *International Conference on Robot & Human Interactive Communication.* IEEE. 65–71.

Michita Imai, Tetsuo Ono, and Hiroshi Ishiguro. 2003. Physical relation and expression: Joint attention for human-robot interaction. *IEEE Transactions on Industrial Electronics* 50, 4, 636–643.

Alexander Birch Jensen. 2021. Towards Verifying a Blocks World for Teams GOAL Agent. *International Conference on Agents and Artificial Intelligence,* (1). 337–344.

Matthew Johnson, Catholijn Jonker, Birna van Riemsdijk, Paul J Feltovich, and Jeffrey M Bradshaw. 2009. Joint activity testbed: Blocks world for teams (BW4T). *International Workshop on Engineering Societies in the Agents World.* Springer. 254–256.

Marcel Adam Just and Patricia A Carpenter. 1976. Eye fixations and cognitive processes. *Journal of Cognitive Psychology* 8, 4, 441–480.

Frederic Kaplan and Verena V Hafner. 2006. The challenges of joint attention. *Journal of Interaction Studies* 7, 2, 135–169.

Moritz Kassner, William Patera, and Andreas Bulling. 2014. Pupil: An Open Source Platform for Pervasive Eye Tracking and Mobile Gaze-based Interaction. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing.* ACM. 1151–1160. https://doi.org/10.1145/2638728.2641695

Ari Kolbeinsson, Erik Lagerstedt, and Jessica Lindblom. 2019. Foundation for a classification of collaboration levels for human-robot cooperation in manufacturing. *Production & Manufacturing Research* 7, 1, 448–471.

Nicole C Krämer, Sabrina Eimler, Astrid Von Der Pütten, and Sabine Payr. 2011. Theory of companions: what can theoretical models contribute to applications and understanding of human-robot interaction? *Journal of Applied Artificial Intelligence* 25, 6, 474–502.

Stephen R H Langton, Roger J Watt, and Vicki Bruce. 2000. Do the eyes have it? Cues to the direction of social attention. *Trends in Cognitive Sciences* 4, 2, 50–59.

Steven M LaValle. 2006. *Planning algorithms.* Cambridge University Press. https://doi.org/10.1017/CBO9780511546877

Hagen Lehmann, Ingo Keller, Reza Ahmadzadeh, and Frank Broz. 2017. Naturalistic conversational gaze control for humanoid robots-a first step. *International Conference on Social Robotics.* Springer. 526–535.

Vladimir Lifschitz. 1987. On the semantics of STRIPS. In *Reasoning about Actions and Plans: Proceedings of the 1986 Workshop,* Michael P Georgeff and Amy L Lansky (Eds.), 1–9. Morgan Kaufmann Publishers.

Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, Wan-Teh Chang, Wei Hua, Manfred Georg, Matthias Grundmann. 2019. Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172.*

Reuth Mirsky, Xuesu Xiao, Justin Hart, and Peter Stone. 2021. Prevention and Resolution of Conflicts in Social Navigation – a Survey. *arXiv preprint arXiv:2106.12113.*

A Jung Moon, Daniel M Troniak, Brian Gleeson, Matthew K X J Pan, Minhua Zheng, Benjamin A Blumer, Karon MacLean, and Elizabeth A Croft. 2014. Meet me where i'm gaz-

ing: how shared attention gaze affects human-robot handover timing. In *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction.* 334–341.

Peter Mundy and Lisa Newell. 2007. Attention, joint attention, and social cognition. *Current directions in psychological science* 16, 5, 269–274.

Jeff Pelz, Mary Hayhoe, and Russ Loeber. 2001. The coordination of eye, head, and hand movements in a natural task. *Experimental Brain Research* 139, 3, 266–277.

André Pereira, Catharine Oertel, Leonor Fermoselle, Joe Mendelson, and Joakim Gustafson. 2019. Responsive joint attention in human-robot interaction. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS).* IEEE. 1080–1087.

David Premack and Guy Woodruff. 1978. Does the chimpanzee have a theory of mind? *Journal of Behavioral and Brain Sciences* 1, 4, 515–526.

Miquel Ramírez and Hector Geffner. 2009. Plan recognition as planning. *Twenty-First International Joint Conference on Artificial Intelligence.*

Miguel Ramírez and Hector Geffner. 2010. Probabilistic plan recognition using off-the-shelf classical planners. *Twenty-Fourth AAAI Conference on Artificial Intelligence.*

Michael Scaife and Jerome S Bruner. 1975. The capacity for joint visual attention in the infant. *Nature* 253, 5489, 265–266.

Ruth Schulz, Philipp Kratzer, and Marc Toussaint. 2018. Preferred interaction styles for human-robot collaboration vary over tasks with different action types. *Frontiers in Neurorobotics* 12, 36.

Shirin Sohrabi, Anton V Riabov, and Octavian Udrea. 2016. Plan Recognition as Planning Revisited. *International Joint Conference on Artificial Intelligence,* 3258–3264.

Leila Takayama, Doug Dooley, and Wendy Ju. 2011. Expressing thought: improving robot readability with animation principles. *ACM/IEEE International Conference on Human-Robot Interaction,* 69– 76.

Michael Tomasello.1995. Joint attention as social cognition. In *Joint attention: Its origins and role in development,* C Moore & P J Dunham (Eds.), (pp. 103–130). Lawrence Erlbaum Associates, Inc.

Michael Tomasello, Malinda Carpenter, Josep Call, Tanya Behne, and Henrike Moll. 2005. Understanding and sharing intentions: The origins of cultural cognition. *Journal of Behavioral and Brain Sciences* 28, 5, 675–691.

Natalie Webb and Tony Renshaw. 2008. Eyetracking in HCI. In *Research Methods for Human-Computer Interaction*, P. Cairns and A. Cox (Eds.) (pp. 35-69). Cambridge: Cambridge University Press. doi:10.1017/CBO9780511814570.004

Robot Learning from Humans in Everyday Life Scenarios

Matthias Hirschmanner, Markus Vincze

Abstract

Robots need to be able to learn about novel environments and acquire new capabilities during deployment. Robot learning from humans is a paradigm to enable the human user to teach robots certain information and skills without programming knowledge. In this chapter, we provide an overview of this domain and present some of our work as concrete examples. First, we address grounded language learning with the goal to create connections between words and references (e.g., objects, locations) in social environments. We present our incremental word learning systems using the Pepper robot. Following that, we introduce to learning low-level actions from demonstrations. We present our systems with an industrial robotic arm and a dexterous robotic hand. Then, we address the role of the teacher in the learning process. We investigate the human factors that are important for facilitating the learning process and present the results of our user studies. We conclude with open challenges and opportunities for further research.

Keywords

Robot Learning, Human-Robot Interaction, Learning from Demonstration, Grounded Language Learning

1 Introduction

Robots are increasingly being placed in unconstrained environments, such as homes, where they must adapt to new situations. They cannot be preprogrammed to perform every task with every object in every environment. They need to be able to learn about new tasks with unseen objects in novel environments. Learning from users' input is one way to acquire this knowledge. Examples of information provided by the user could include demonstrating a task or providing language feedback via speech. Robotic learning from humans enables novice users to teach new tasks to a robot without extensive programming knowledge. Therefore, the topic of learning from human teachers has received increased attention in recent years [Ravichandar et al. 2020].

Chernova and Thomaz [2014] motivate learning from humans using robots in the household. Vacuum cleaning robots have become ubiquitous in recent years. They can be placed in an unknown environment and start operating immediately. They can even create a map of the environment to navigate from room to room autonomously. This works well as long as certain constraints are met, such as a flat floor without stairs, cables, or other obstacles.

A general-purpose household robot must complete a much wider and more complex set of tasks. A user would expect it to empty the dishwasher, clean the bathtub, or store objects in their designated storage location. These tasks are not only more complex in terms of manipulation and perception but also need to be performed in less constrained environments. Each household is unique and

https://doi.org/10.34727/2022/isbn.978-3-85448-052-5_8

different from other households. There could be similarities that can be exploited, such as the same type of existing object (e.g., cupboards, drawers, or fridges) or the same type of room (e.g., kitchen, bathroom). However, the storage location of certain objects, such as plates, mugs, or cookie jars, can be unique and arbitrary for each home. These conditions cannot be preprogrammed into the robotic knowledge base in the factory but must be learned by a robot, once it arrives in a new household, similar to a new person moving in. There has to be the possibility for the user to extend the robot's knowledge and modify its behavior. Learning from demonstrations (LfD) methods attempt to learn information, and action policies (i.e., how to perform a task) from examples provided by humans.

Additionally, household robots must be controlled by users directly. A popular and intuitive approach is to use voice commands such as *"Put the strawberry jam into the food storage cabinet."* Modern speech recognition algorithms perform well and can convert a spoken language to text even from a distance, as demonstrated by stationary voice assistants integrated into speakers at users' homes [Berdasco et al. 2019]. A more challenging task is to make sense of what has been said. A robot might not know which object is meant by *"strawberry jam"*, what location by *"food storage cabinet"* and maybe not even how to perform the action *"put"*. "Grounded Language Learning" [Matuszek 2018] is the process of assigning words to references in physical and social spaces. It is a subfield of robotic learning from humans but is often not mentioned in the context of LfD.

Human factors are an important consideration when learning from human teachers. Many papers focus on algorithms for learning policies from demonstrations. The role of human teachers is often overlooked. Especially, novice users cannot be treated as infallible oracles who always provide perfect demonstrations to the robot. Instead, users are part of the learning loop and influence the final performance of the robot immensely. A learning system must consider the human in the loop and accommodate their needs.

The field of "Robotic learning from humans" is very broad, with many different application fields. However, we focus on two domains as an example to provide a starting point for discussing the human factors connected to the learning system. The main contributions of this chapter are:

- We give an introduction to the field of grounded language learning and present our framework with the Pepper robot. It is focused on iterative language learning and being transparent towards the human teacher.

- We discuss the topic of learning low-level actions from human demonstrations and give an overview of recent approaches. We present our setups with an industrial robotic arm and a dexterous robotic hand in simulation.

- We investigate different human factors involved in the learning process such as the teacher's workload, self-efficacy, transparency and trust. We present the results of our experiments with a robot teleoperation setup, a language-learning setup and an interaction scenario.

In Section 2, we provide a brief overview of the field of "grounded language learning," highlighting our two approaches with the Pepper robot. Section 3 discusses "learning of low-level actions" with examples using industrial robotic arms and simulated robotic hands for dexterous manipulation, with a special emphasis on input methods. In Section 4, we discuss the teacher side of the learning loop to identify human factors that must be considered when building learning systems, such as workload, self-efficacy, and trust. Section 5 concludes the paper and mentions opportunities for further research in this field.

2 Grounded Language Learning

Robots are increasingly being used in environments where they must be controlled by untrained nonexpert users. Using one's voice to give commands or communicate intent is a very natural approach in everyday life. Therefore, speech is a very popular modality for giving instructions to robots and has been extensively studied [Matuszek 2018].

Grounded language (also known as situated language) connects the natural language to references in physical and social spaces [Tellex et al. 2020]. For example, the word *"mug"* can be connected to a class of objects, the word *"fridge"* to a storage location different for each home, or the word *"put"* with a series of motor controls dependent on the specific object. The purpose of grounded language learning is to create these word-reference connections.

Many datasets have been introduced because of the various scenarios to which grounded language learning can be applied. An early example was the MARCO dataset [MacMahon et al. 2006], which addresses the problem of navigation instructions. It consists of navigation instructions for a simulated robot (e.g., *"With the wall on your left, walk forward."*). The goal of the system is to understand and follow these instructions with a simulated robot. Other examples of datasets that can be used as starting points for language learning are object detection datasets. They provide natural language class labels for the images. Imagenet has many class labels (e.g., *snail*, *broccoli*, *teapot*) [Deng et al. 2009]. It uses the WordNet [Fellbaum 1998] hierarchy of sets of synonyms that describe meaningful concepts by adding images to each set. Other datasets extend image labels to describe en-

tire images, such as *"a kid sitting on the side walk eating a slice of pizza."* in the COCO dataset [Lin et al. 2014].

Robots have a various sensors that enable them to use different modalities for grounded language learning such as detected objects, human movements, and recognized actions. Multimodal datasets are used to cover more modalities of real-life scenarios than the above-mentioned. Gaspers et al. [2014] present a dataset where human participants show object manipulation actions to a robot and explain what they are doing. It includes video, audio and human posture data.

We introduced the action verb corpus dataset geared toward object manipulations [Gross et al. 2018], consisting of 390 simple actions (i.e., *take*, *put*, and *push*) of 12 humans following pictured instructions of tasks and describing what they are doing. It includes audio, video and motion data of hand joints and objects. The dataset is annotated with utterance transcriptions, part-of-speech tags, which object is currently moved, and whether a hand touches an object, or an object touches the ground/table. This type of cross-modal and cross-situational data can be used to create systems that learn from humans demonstrating actions while explaining what they are doing. A robot could infer the object name of a manipulated object and the name of the action. The action could be defined by its outcome or by its trajectory. The data can also be used by the robot to replicate the presented action.

Cooccurrence statistics of words and references are often used in computational models that learn from this type of cross-situational data [Krenn et al. 2020]. Taniguchi et al. [2017] provided an overview of different approaches. However, these methods often require large datasets or batches of examples for learning, which is often disadvantageous when deploying a robot in a new environment to learn about new concepts from a human teacher. Additionally, noisy real-world data collected by a robot usually differ from those provided on datasets. Consider a situation where to teach a new concept to a robot, the user must first gather a dataset, which is of course cumbersome and not feasible for a robot at home. However, an incremental learning system uses each new sample to update the probability of a word-reference pair.

We introduced a word-learning system for the Pepper[1] robot, as a concrete example, in Hirschmanner et al. [2018a]. The goal is to learn word-object and word-action mappings in a human-robot interaction scenario. The setup and system architecture are shown in Figure 1. The human teacher demonstrates actions (i.e., *take*, *put*, *push*) to the robot and explains what they are doing. The system infers the type of action from the movement of an object obtained from an object detector processing visual data. The output of the speech recognition module

1 https://www.softbankrobotics.com/emea/en/pepper

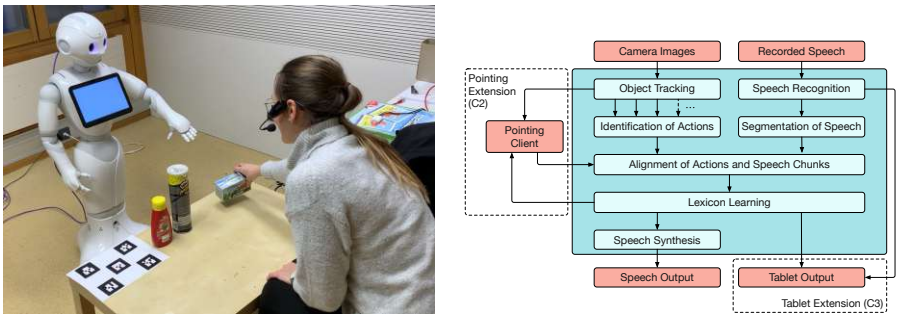


Figure 1 A user performs an object manipulation action (left). Overview of the system architecture (right) used for the language learning system. Tracked objects are used to identify actions and then aligned with the utterances. Normalized pointwise mutual information is used to estimate object/action-word cooccurrence. From Hirschmanner et al. [2021].

is aligned with the action to create utterance-situation pairs. An example of an episode with two utterance-situation pairs would be <*I take the box - ACTION1 OBJECT1*>, <*and put it next to the can. - ACTION2 OBJECT1 OBJECT2*>. We use the normalized pointwise-mutual information (*npmi*), which is a measure of the likelihood of an object/action-word cooccurrence. The *npmi* value is updated after each detected situation-utterance pair. We propose two extensions to this system to increase transparency for the human teacher, in Hirschmanner et al. [2021]. These extensions will be addressed in more detail in Section 4.

The approaches described above usually treat the robot as a passive observer. However, unlike a computer program, a robot is an embodied agent, which can actively request new information by directing the attention of the human teachers toward some unknown references through pointing, gaze, or verbal utterances. This can also be motivated by findings in the developmental psychology of children during language acquisition. They actively request the names of objects using deictic gestures, such as pointing or gaze [Krenn et al. 2019]. A robot can formulate full sentences to acquire knowledge of its surrounding. At public events, we experimented with a Pepper robot that points at objects and formulates questions about the objects pointed at [Hirschmanner et al. 2018b]. The questions did not only refer to the name of an object (i.e., *"How do you call this object?"*) but also to its function (i.e., *"What do you use it for?"*) and the users' preferences (i.e., *"How do you like it?" "What does it mean to you?"*). We used a relatively simple approach that uses part-of-speech tagging to identify nouns, verbs, and adjectives in users' responses. The number of occurrences of each word in these categories is summed up for each object, providing the robot information about the objects modeled using the cultural space model [Schürer et al. 2018]. In this

preliminary study, we looked at how human teachers respond to questions from robots.

This section provides a brief introduction to grounded language learning. We want to motivate further development of incremental and active word learning systems for robots, similar to Bisk et al. [2020]. For a general introduction to robots that use language, we refer to Tellex et al. [2020].

3 Learning Low-Level Actions

When deploying a service robot at home, it can already perform certain actions, such as grasping objects and placing them somewhere. In our example of the household robot, the user might give the voice command *"Put the salad bowl into the dishwasher."* Assume that it has already learned which object is meant by *"salad bowl"* and which location by *"dishwasher"* through grounded language learning. There could be a problem in which the robot puts the bowl into an unsatisfactory position or is unable to place the bowl at all. The user will probably know a good strategy for positioning the bowl in the dishwasher. The user can teach the robot the low-level action of placing this specific bowl into the dishwasher using an LfD algorithm.

When creating an LfD system, the following numerous design decisions must be addressed. Which input method is used by the teacher to demonstrate the action? How is the demonstration represented (i.e., which state space is used)? Which algorithm is used to learn the presented demonstration? We give a short overview of the different possibilities to address these design decisions. At the end of the section, we present some concrete projects where we implement learning action policies from human demonstration. We direct the interested readers to Billard et al. [2016] and Chernova and Thomaz [2014] for a general introduction to the topic. A detailed view of the algorithms used in LfD can be found in Osa et al. [2018]. Recent advances are summarized in Ravichandar et al. [2020].

A human teacher can provide demonstrations to a robot in several different ways. Teleoperation is a popular method. The human teacher controls the robot via some device, such as a keyboard, mouse, or joystick to make the robot directly perform the action that is to be learned, which is often cumbersome and difficult to do for novice users [Whitney et al. 2020]. To overcome these limitations, researchers investigated using methods, such as motion tracking to replicate the human motion on the robot [Chernova and Thomaz 2014]. Kinesthetic teaching is an alternative to teleoperation, in which a human manually guides the end-effector of the robot to perform the task [Ravichandar et al. 2020]. For tele-

operation or kinesthetic teaching, the sensor data of the robot (e.g., joint angles, end-effector positions, and torques) can be recorded directly and used as input for the machine learning algorithm. We compare kinesthetic teaching to teleoperation on a Pepper robot concerning to the workload on the human teacher in Hirschmanner et al. [2019], which is summarized in Section 4.

Alternatively, some approaches exist that learn directly from observing a human performing an action, making teaching much easier and more natural to the human teacher. The drawback the machine learning problem becomes more difficult because the human movements must be encoded or mapped to the robot's movement [Ravichandar et al. 2020]. Other technical problems may occur if the human performs the task in a way that the robot cannot properly perceive (e.g., fast movements, occlusions, or leaving the field of view).

The next design decision is how to store and process the demonstrations. In this chapter, we will mainly discuss deriving a policy $\pi : \mathcal{S} \rightarrow \mathcal{A}$ that maps from a state vector $\boldsymbol{s} \in \mathcal{S}$ to a low-level action $\boldsymbol{a} \in \mathcal{A}$. Other approaches learn policies that output complete trajectories instead of low-level actions. Instead of policies, alternative learning outcomes in LfD can be plans or a reward function for reinforcement learning (i.e., Inverse Reinforcement Learning) [Ravichandar et al. 2020]. The choice of state space $\mathcal{S}$ and action space $\mathcal{A}$ depends on the concrete problem statement. A very simple state space $\mathcal{S}$ may represent the current time, resulting in an open-loop control, where no feedback on the robot or its environment is provided to the policy. Additionally, the robot's sensor data, such as end-effector positions, joint angles, joint velocities, and torques can be used. Sensor data from the environment of the robot can also be included, which can be high-level, such as the pose of an object received by an object pose estimator or low-level, such as a light detection and ranging sensor (LIDAR) or raw camera images.

Similarly, the action space $\mathcal{A}$ can be defined in different ways. Low-level policies could output torques applied to each robot joint. Motion controllers can be used to output actions as end-effector poses or velocities in Cartesian or joint space. Actions can also be defined as trajectories or even sub-tasks as a high-level representation. The choice of granularity of the state and action space depends on the concrete problem, as previously stated. Naturally, the state depends on the available sensors and the teaching approach. For example, when using kinesthetic teaching, using raw camera images as the input might be problematic because the human teacher moving the robot is only present during the demonstration phase. Thus, the image would necessarily be different when the robot executes the action without a teacher. Additionally, a balance should be found between providing enough information to accurately represent the demonstration and not introducing too many dimensions to make the machine learning problem too dif-

ficult (“curse of dimensionality” [Bellman 1957]). Similarly, for the action space, a simple representation that can still perform the required task is preferable. For example, for a task involving pushing an object on a table, the two-dimensional (2D) position of the end-effector at a fixed distance to the table might be sufficient. If a device is used to teleoperate the robot, the obvious choice for the action space would be the same domain that is used by the demonstrator, such as the steering angle and acceleration for a remote-controlled car.

A recorded trajectory $\boldsymbol{\tau}$ consists of a state vector $\boldsymbol{s}$ and an action vector $\boldsymbol{a}$ per timestep. The complete demonstration $\mathcal{D}$ can then be defined as

$$\boldsymbol{\tau} = [\boldsymbol{s}_0, \boldsymbol{a}_0, \boldsymbol{s}_1, \boldsymbol{a}_1, \dots \boldsymbol{s}_{T-1}, \boldsymbol{a}_{T-1}, \boldsymbol{s}_T, \boldsymbol{a}_T], \quad \mathcal{D} = \{\boldsymbol{\tau}_i\}_{i=1}^{N}.$$

Training a policy $\pi(\boldsymbol{s}) = \boldsymbol{a}$ from these demonstrations can be seen as a supervised learning problem. Over the years, many different supervised learning approaches have been applied to LfD. Popular approaches include support vector machines (SVM) [Chernova and Veloso 2009], Gaussian mixture models (GMMs) [Khansari-Zadeh and Billard 2011], and Gaussian processes [Choi et al. 2016]. In recent years, artificial neural networks (ANNs) have gained popularity (e.g., Rahmatizadeh et al. [2018]; Zhang et al. [2018]; Young et al. [2021]). There have also been many approaches that address specific problems occurring in LfD. For example, the DAGGER algorithm reduces the number of demonstrations required and, therefore, the load on the human teacher by generating additional demonstrations [Ross et al. 2011].

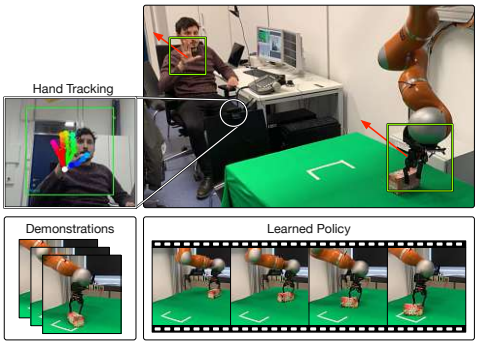


Figure 2 A user teleoperating a Kuka robotic arm using hand tracking to perform a task. The demonstrations are used to learn a policy represented as a neural network. From Hirschmanner et al. [2020].

We present an LfD approach in Hirschmanner et al. [2020], as a concrete example of how to address the different design decisions. We trained a policy on the Kuka LWR IV+ [Bischoff et al. 2010] robot to push a box to a certain position on the

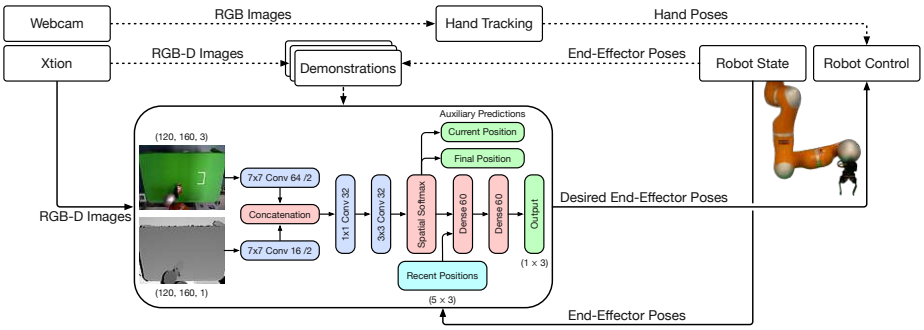


Figure 3 Overview of the system. The dashed lines represent the procedure for collecting demonstrations for training. The continuous lines represent the information flow during policy execution. From Hirschmanner et al. [2020].

table. The demonstrations were recorded using a teleoperation setup based on hand tracking from an RGB webcam. The setup is shown in Figure 2. The state of the robot and its environment are represented as an RGB-D image and the end-effector position of the robot in Cartesian space at the five previous timesteps. For the actions, we use the relative end-effector position $\Delta \boldsymbol{p} \in \mathbb{R}^3$ in Cartesian space. These representations were chosen to capture the entire scene without requiring a separate method to obtain the object pose. The policy is represented as a convolutional neural network (CNN) based on the architecture of Zhang et al. [2018]. It includes two auxiliary tasks during the policy training to predict the current and final end-effector position from the input images. The architecture is shown in Figure 3. We recorded 98 demonstrations at a 10 Hz sampling rate. For the evaluation, we placed the box in different positions on the table, which were unseen during the demonstrations. The robot started to push the box in 86.1% of the trials and reached the goal in 58.3%.

These results indicate some problems with pure supervised learning methods. Demonstrations will not cover each possible configuration in the problem space. During the policy execution, the agent encounters situations unseen during the demonstrations. The situation when the source domain distribution differs from the target domain distribution is referred to as a “covariate shift” [Osa et al. 2018]. Several data-efficient trajectory-learning methods addressed this generalization problem recently. Task-parametrized models of movement [Calinon 2016] use GMMs and represent demonstrations in different frames of reference to improve generalization. Probabilistic movement primitives [Paraschos et al. 2018] represent movement policies in the form of a distribution of trajectories that can be conditioned on desired via-points to adapt to new situations. Kernelized movement primitives [Huang et al. 2019] extend this idea to a nonparametric approach

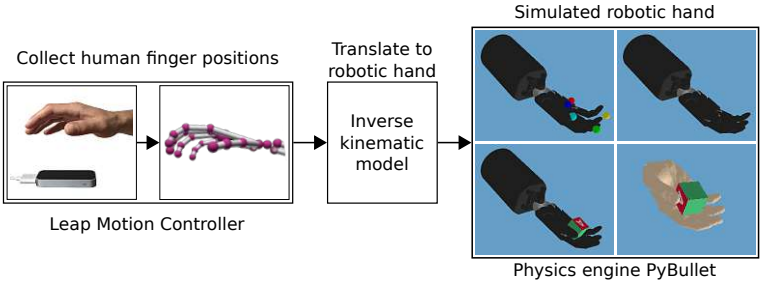


Figure 4 Teleoperation system used to collect the dexterous manipulation tasks. The Leap Motion hand tracker is used to control a simulated robotic or human hand in simulation. The two tasks are shown in the left column of the image on the right. From Zahlner et al. [2020].

geared toward high-dimensional inputs and extrapolation of demonstrated trajectories. One limitation of these trajectory-learning approaches is that the task parameters, such as object poses or obstacles, must be provided to the system when executing the policy, for example, by computer vision algorithms [Pervez et al. 2017]. Additionally, they require a motion planner that converts the trajectory to low-level actions.

One problem with supervised learning is that a learned policy will not outperform the teacher. Researchers have worked on using expert demonstrations in reinforcement learning, as an alternative. In this alternative learning paradigm, the agent can discover new policies through exploration. A reward function is required, which returns a value depending on how beneficial a certain step is to achieve the goal of the task. The machine learning algorithm attempts to maximize the sum of rewards over all timesteps. In the previous box pushing example, this reward function could be the negative distance of the box from the goal. Reinforcement learning is usually very time-intensive because actions that solve a certain task must be discovered through exploration. When expert demonstrations that solve the task are available, this process can be speed-up (e.g., Nair et al. [2020]).

Similarly, we used demonstrations to accelerate the learning process for two dexterous manipulation tasks in Zahlner et al. [2020]. The setup consists of the Shadow Dexterous Hand[2] in the PyBullet[3] simulator. The tasks involved reaching a target position for each fingertip and manipulation of a block to rotate it to a certain orientation. Demonstrations are provided using a teleoperation system that tracks the human hand using a Leap Motion Controller[4] to replicate the current

2 https://www.shadowrobot.com/dexterous-hand-series/

3 https://pybullet.org

hand poses on the simulated hand. The teleoperation system and the different tasks are shown in Figure 4. The state space consists of the absolute angle and velocity of all 20 joints and additional task-specific data. For the reaching task, the current and target Cartesian positions of the fingertips are added to the state. For the object manipulation task, the cube's current and target Cartesian poses, as well as its linear and angular velocities, are provided. The action space of both tasks consists of the 20-dimensional noncoupled hand joints. Both tasks were designed to be similar to the ones presented by Plappert et al. [2018]. We trained the policy with deep deterministic policy gradient (DDPG) [Lillicrap et al. 2016] and hindsight experience replay (HER) [Andrychowicz et al. 2017]. The policy was represented as a neural network. We used demonstrations for pre-training the policy using supervised learning. We saw a speed-up compared to reinforcement learning without pre-training from $2.2 \cdot 10^6$ to $1.2 \cdot 10^6$ timesteps for the reaching task. No comparable speed-up was observed for the cube manipulation task. We hypothesize that this is because the goal in the manipulation task is often reached randomly during exploration and thus does not profit from demonstrations. Additionally, the quality of the demonstrations was low because of the difficulty in manipulating a cube in the simulation without haptic feedback.

Learning the reward function from expert demonstrations is another approach to combining demonstrations and reinforcement learning. This domain is known as inverse reinforcement learning (IRL). The main idea is that the teacher performs demonstrations that optimize an unknown reward function. IRL approaches try to find this reward function. This problem is ill-posed since the expert's behavior could be explained using multiple functions. The retrieved reward function is then used to train a motion policy using standard reinforcement learning algorithms in a subsequent step. Because of to the limited scope of this chapter, we refer to Osa et al. [2018] and Arora and Doshi [2021] for an extensive overview of IRL.

4 Human Factors

In the previous sections, we have addressed how a robot can use the information provided by a human teacher to acquire new skills and knowledge. We did not discuss the influence of the learning process on the user and vice-versa. The human teacher is a part of the learning loop and significantly affects the final performance of the robot. A learning system must consider the humans in the loop and accommodate their needs. However, few studies have been conducted to evaluate the role of the teacher and how the teaching behavior influences a learn-

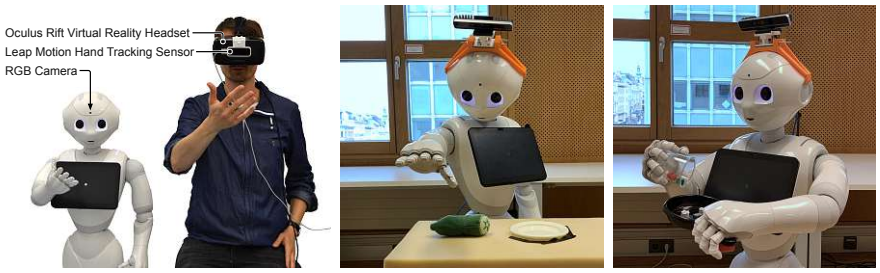
4 https://www.leapmotion.com

ing system [Sena and Howard 2020]. See Vollmer and Schillingmann [2018] for a recent review.

When designing a system that learns from humans, several factors must be considered. The teaching process must be designed in a way to keep the workload of the user minimal. Low mental and physical workloads lead to higher quality and quantity of training data by keeping the human teacher motivated and engaged [Cui et al. 2021]. The quality of the training data directly affects the learning outcome.

In the context of learning low-level actions, we compared the workload of human demonstrators using a virtual-reality teleoperation setup and kinesthetic guidance in Hirschmanner et al. [2019]. The human teacher wears a virtual reality headset with an attached Leap Motion Controller to teleoperate the Pepper robot shown in Figure 5. The camera stream is displayed in the headset. The current head orientation of the user is imitated by the robot. The hand pose is tracked using the Leap Motion Controller, which is also transferred to the robot. Thus, the robot imitates the upper-body movements of the user. The robot's physical dimensions and constraints are different from those of humans. However, humans can still complete the task successfully because they receive immediate feedback and can adapt to the situation. We compared this setup to kinesthetic guidance, in which users moved the arms of the robot manually. In a user experiment (n=21), participants performed an object grasping task and a pouring task that required controlling both of the robot's arms. Most of the users preferred the teleoperation system for both tasks stating because it was easier to learn. The workload was measured using the NASA-TLX questionnaire [Hart and Staveland 1988]. Compared to kinesthetic guidance, the workload of the users was lower when using teleoperation for the pouring task. We also observed a reduction in task duration for the pouring task when using the teleoperation setup, as an objective measure. Contrary to these results, previous research demonstrated that users preferred kinesthetic guidance to teleoperation [Fischer et al. 2016; Praveena et al. 2019]. This is not contradictory; rather, it emphasizes the importance of tailoring the teaching method to the concrete scenario.

Another important factor that contributes to the workload of the teacher is the number of demonstrations required to train an algorithm. Approaches based on deep learning often require many demonstrations to reach satisfactory performance. Mandlekar et al. [2021] report that 40 demonstrations from a proficient teacher were sufficient to train simple actions, such as lifting an object. For a more complex task, such as transporting a hammer from the workspace of one robot arm to the workspace of another robot arm with a handover operation, the success rate dropped from 72% when using 200 demonstrations to 30.7% when using 40 demonstrations. To overcome the sample inefficiency of deep learn-



(a) Teleoperation Setup **(b)** Grasping Task **(c)** Pouring Task

Figure 5 The human demonstrator uses the virtual reality teleoperation setup to control the Pepper robot to perform two different tasks. The human head and hand poses are then transferred to the robot. From Hirschmanner et al. [2019].

ing approaches, one can start with a pretrained policy and only ask for human demonstrations if the robot fails (e.g., DelPreto et al. [2020]) or a policy trained for a different task and apply meta-learning with a low number of demonstrations to transfer it to a new task (e.g., Finn et al. [2017]). Algorithms that learn trajectories instead of low-level actions using GMMs or movement primitives (e.g., Calinon [2016]; Paraschos et al. [2018]; Huang et al. [2019]) are designed to require few demonstrations (<10) but require task-parameters, such as object poses.

Additionally, the teacher's mental model of the learning system should align with the actual model to facilitate good teaching behavior of the user [Cakmak and Thomaz 2014]. A robot needs to be able to communicate the current state of the learning system and how the teacher can improve teaching examples to the teacher. These topics are also investigated in the context of transparency in human-robot interaction and explainable artificial intelligence (AI) to increase trust in robots [Papagni and Koeszegi 2021].

Robots, as embodied agents, can expose the current state of the user through various means, such as visualization, movements, text, speech, lights, and imagery [Wallkötter et al. 2021]. A combination of these different modalities is often used. In Hirschmanner et al. [2021] we investigated the efficacy of different modalities. We integrated transparency mechanisms using visualization and deictic gestures in our word-learning system described in Section 2. As a visualization, the Pepper robot displays its current lexicon and the output of the speech recognition system on its screen. The robot uses deictic gestures, such as looking and pointing at objects to either request additional information or to announce the learned word of the object. This behavior is motivated by early-childhood language learning in humans [Krenn et al. 2019]. We did not observe any significant performance difference between the base, visualization, and deictic gestures con-

ditions in a user experiment (n=32). However, the users' knowledge of the system's state positively correlates with the self-reported perception of control and perceived learning success. Users exhibited more interactive behavior when the robot used deictic gestures which might help keep the user engaged, but it also increases noise in the training data. These results encourage further investigation of the transparency mechanism in LfD systems.

Additionally, a learning system should consider factors that influence the user's self-efficacy and perceived control when teaching the system. Self-efficacy is the confidence of a user in being able to perform actions to accomplish a certain task [Bandura 1982]. In the context of a teaching system, self-efficacy is the confidence in being able to teach a new task or concept to a robot. High self-efficacy is important to increase the user's willingness to engage with a robot and to keep them motivated to interact with a learning system in a long-term deployment [Pütten and Bock 2018; Robinson et al. 2020].

The way the robot interacts with the user can influence these factors. We conducted a user experiment (n=29) in Zafari et al. [2019] to study the effect of the interaction style. The task of the user was to build a house of cards. The Pepper robot observed the user and interacted with them using natural language output, such as *"Very nice, keep up the good work."* The speech output was controlled by a researcher following a script. In the person-oriented condition, the robot used motivational sentences to support the user. In the task-oriented condition, the interaction was focused on the task progress and pushed the participant to improve their performance. In the neutral control condition, the robot was only a game instructor and commented on the task progress. We did not tell the participants that they were demonstrating how to build a house of cards to the robot, but the scenario could be used for an LfD system. We found that users in the person-oriented condition reported higher self-efficacy and that they experienced the interaction as less frustrating than in the task-oriented condition. Additionally, participants performed the task significantly longer and thus stayed engaged for a longer time in the person-oriented conditions than in the neutral condition. These results indicate that the interaction style of a robot can also be used to positively influence the human demonstrator and as a consequence, they might be willing to provide more training data in learning from a human setting.

Another important factor to investigate is how trust is influenced in learning from humans setting. A low trust may cause the human teacher to abandon the system. Over-trusting the system may lead to the user ending the teaching process before the system has learned a task reliably and failing to monitor the trained agent, which may result in unwanted behavior or even accidents [Lewis et al. 2018]. DelPreto et al. [2020] found that low accuracy in an LfD task reduces trust and increases the users' workload. They also found tendencies that users

overestimate the robot's skills. Hedlund et al. [2021] found that when robots fail to perform the learned tasks, participants' trust in the robot and themselves as teachers decreases.

5 Conclusion and Open Challenges

In this chapter, we presented our work and set it into the context of the field of robotic learning from humans. First, we motivated the need for grounded language learning of social robots, i.e., connecting words with references such as objects. Learning these connections is required for a robot to follow voice commands. We presented two word-learning systems using the Pepper robot. Following that, we addressed the field of learning low-level actions from demonstrations. We covered the main design choices that must be made when developing a learning system. We presented two systems with different robotic setups to demonstrate different design decisions. Furthermore, we discussed the role of the human teacher in the learning system. We emphasized the importance of considering factors, such as workload, self-efficacy, and trust during the teaching process to obtain good training examples and keep the user motivated. We presented three user studies for different robotic setups that investigated workload, transparency, and self-efficacy.

The field of learning from human users is emerging, as more robots move into living spaces. There are still many open problems to be tackled. Robots need to be able to acquire information from spoken language to make interactions with humans more natural. Grounded language learning methods that can incrementally process the high-dimensional multimodal data that robots will encounter in everyday situations must be developed [Bisk et al. 2020]. Additionally to spoken language, they need to be able to understand nonverbal communication to better interact with humans.

Learning action policies from demonstrations has accelerated in recent years [Ravichandar et al. 2020]. Many algorithms have been developed to address the special conditions and constraints associated with learning from human teachers. However, it is often difficult to compare the approaches because of the limited number of available benchmarks using real demonstrations provided by humans that have advanced other fields such as computer vision or reinforcement learning. Two of the few examples are Mandlekar et al. [2021] and [Sharma et al. 2018]. New standardized benchmarking methods on real robotic systems will be required to advance the field of learning motion policies from human demonstrations.

Demonstrations are usually task-specific and do not cover the entire problem space. A promising direction for further research is to develop algorithms that

generalize better across tasks, domains, and robots. A combination of demonstrations with reinforcement learning could be useful in this regard and should be examined further. Demonstrations can be used to shorten the long training times of reinforcement learning algorithms. Additionally, these systems often require tedious hyperparameter tuning, which is not feasible for novice users. Further research is required to develop methods that require few hyperparameters and are easy to tune automatically.

The role of the teacher and teaching behavior have been under-represented in the robotic learning from humans pipeline [Vollmer and Schillingmann 2018]. High-quality training data from the human teacher facilitates the learning process. To ensure this, the teacher must be considered as a part of the learning loop when designing a system. Further research should aim to create non-intrusive and intuitive teaching systems to minimize the workload of the user and keep them motivated and engaged.

If we want to deploy robots that learn from humans in users' homes, the effect of the learning system on the users must be studied further. Users will only accept these systems, if they see an added value in them and if they enjoy using them [de Graaf et al. 2017]. We believe that self-efficacy is an important concept in that regard. We must investigate which factors influence the trust of the user in the system to find a balance between not overtrusting the system and trusting it enough to use it continuously.

Bibliography

Marcin Andrychowicz, Filip Wolski, Alex Ray, Jonas Schneider, Rachel Fong, Peter Welinder, Bob McGrew, Josh Tobin, OpenAI Pieter Abbeel, and Wojciech Zaremba. 2017. Hindsight experience replay. In *Advances in neural information processing systems*. 5048–5058.

Saurabh Arora and Prashant Doshi. 2021. A survey of inverse reinforcement learning: Challenges, methods and progress. *Artificial Intelligence* 297 (2021), 103500. https://doi.org/10.1016/j.artint.2021.103500

Albert Bandura. 1982. Self-efficacy mechanism in human agency. *American psychologist* 37, 2 (1982), 122. Publisher: American Psychological Association.

Richard Bellman. 1957. *Dynamic Programming* (1 ed.). Princeton University Press, Princeton, NJ, USA.

Ana Berdasco, Gustavo López, Ignacio Diaz, Luis Quesada, and Luis A Guerrero. 2019. User experience comparison of intelligent personal assistants: Alexa, Google Assistant, Siri and Cortana. *Multidisciplinary Digital Publishing Institute Proceedings* 31, 1 (2019), 51.

Aude G Billard, Sylvain Calinon, and Rüdiger Dillmann. 2016. Learning from Humans. In *Springer Handbook of Robotics*, Bruno Siciliano and Oussama Khatib (Eds.). Springer International Publishing, Cham, Chapter 74, 1995–2014.

Rainer Bischoff, Johannes Kurth, Günter Schreiber, Ralf Koeppe, Alin Albu-Schäffer, Alexander Beyer, Oliver Eiberger, Sami Haddadin, Andreas Stemmer, Gerhard Grunwald and Gerhard Hirzinger. 2010. The KUKA-DLR Lightweight Robot arm-a new reference platform for robotics research and manufacturing. In *ISR 2010 (41st international symposium on robotics) and ROBOTIK 2010 (6th German conference on robotics)*. VDE, 1–8.

Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, Nicolas Pinto, and Joseph Turian. 2020. Experience Grounds Language. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Online, 8718–8735. https://doi.org/10.18653/v1/2020.emnlp-main.703

Maya Cakmak and Andrea L Thomaz. 2014. Eliciting good teaching from humans for machine learners. *Artificial Intelligence* 217 (2014), 198–215.

Sylvain Calinon. 2016. A tutorial on task-parameterized movement learning and retrieval. *Intelligent Service Robotics* 9, 1 (Jan. 2016), 1–29. https://doi.org/10.1007/s11370-015-0187-9

Sonia Chernova and Andrea L Thomaz. 2014. *Robot learning from human teachers.* Morgan & Claypool Publishers.

Sonia Chernova and Manuela Veloso. 2009. Interactive policy learning through confidence-based autonomy. *Journal of Artificial Intelligence Research* 34 (2009), 1–25.

Sungjoon Choi, Kyungjae Lee, and Songhwai Oh. 2016. Robust learning from demonstration using leveraged Gaussian processes and sparse-constrained optimization. In *2016 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 470–475.

Yuchen Cui, Pallavi Koppol, Henny Admoni, Scott Niekum, Reid Simmons, Aaron Steinfeld, and Tesca Fitzgerald. 2021. Understanding the Relationship between Interactions and Outcomes in Human-in-the-Loop Machine Learning. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence.* International Joint Conferences on Artificial Intelligence Organization, Montreal, Canada, 4382–4391. https://doi.org/10.24963/ijcai.2021/599

Maartje de Graaf, Somaya Ben Allouch, and Jan van Dijk. 2017. Why Do They Refuse to Use My Robot?: Reasons for Non-Use Derived from a Long-Term Home Study. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction. ACM, Vienna Austria*, 224–233. https://doi.org/10.1145/2909824.3020236

Joseph DelPreto, Jeffrey I Lipton, Lindsay Sanneman, Aidan J Fay, Christopher Fourie, Changhyun Choi, and Daniela Rus. 2020. Helping Robots Learn: A Human-Robot Master-Apprentice Model Using Demonstrations via Virtual Reality Teleoperation. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 10226–10233.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*. IEEE, 248–255.

Christiane Fellbaum (Ed.). 1998. *WordNet: An Electronic Lexical Database.* MIT Press, Cambridge, MA.

Chelsea Finn, Tianhe Yu, Tianhao Zhang, Pieter Abbeel, and Sergey Levine. 2017. One-shot visual imitation learning via meta-learning. In *Conference on robot learning.* PMLR, 357–368.

Kerstin Fischer, Franziska Kirstein, Lars Christian Jensen, Norbert Krüger, Kamil Kukliński, Maria Vanessa aus der Wieschen, and Thiusius Rajeeth Savarimuthu. 2016. A Comparison of Types of Robot Control for Programming by Demonstration. In *The Eleventh ACM/IEEE International Conference on Human Robot Interaction (HRI '16)*. IEEE Press, 213–220. http://dl.acm.org/citation.cfm?id=2906831.2906868

Judith Gaspers, Maximilian Panzner, Andre Lemme, Philipp Cimiano, Katharina J Rohlfing, and Sebastian Wrede. 2014. A multimodal corpus for the evaluation of computational models for (grounded) language acquisition. In *Proceedings of the 5th Workshop on Cognitive Aspects of Computational Language Learning (CogACLL)*. 30–37.

Stephanie Gross, Matthias Hirschmanner, Brigitte Krenn, Friedrich Neubarth, and Michael Zillich. 2018. Action Verb Corpus. In *Proc. 2018 Language Recognition and Evaluation Conference.* Miyazaki, Japan.

Sandra G Hart and Lowell E Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. *Advances in psychology* 52 (1988), 139–183.

Erin Hedlund, Michael Johnson, and Matthew Gombolay. 2021. The Effects of a Robot's Performance on Human Teachers for Learning from Demonstration Tasks. In *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction.* 207–215.

Matthias Hirschmanner, Stephanie Gross, Brigitte Krenn, Friedrich Neubarth, Martin Trapp, and Markus Vincze. 2018a. Grounded Word Learning on a Pepper Robot. In *Proceedings of the 18th International Conference on Intelligent Virtual Agents*. ACM, 351–352.

Matthias Hirschmanner, Stephanie Gross, Setareh Zafari, Brigitte Krenn, Friedrich Neubarth, and Markus Vincze. 2021. Investigating Transparency Methods in a Robot Word-Learning System and Their Effects on Human Teaching Behaviors. In *2021 30th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN).* Vancouver, BC, CA (Virtual Conference), 175–182.

Matthias Hirschmanner, Ali Jamadi, Bernhard Neuberger, Tim Patten, and Markus Vincze. 2020. Learning Manipulation Tasks from Vision-based Teleoperation. In *Proceedings of Joint Austrian Computer Vision and Robotics Workshop (ACVRW2020)*. 42–47. https://doi.org/DOI:10.3217/ 978-3-85125-752-6-11

Matthias Hirschmanner, Oliver Schürer, Brigitte Krenn, Christoph Müller, Friedrich Neubarth, and Markus Vincze. 2018b. Improving the Quality of Dialogues with Robots for Learning of Object Meaning. In *Workshop on Language and Robotics at IROS2018*.

Matthias Hirschmanner, Christiana Tsiourti, Tim Patten, and Markus Vincze. 2019. Virtual Reality Teleoperation of a Humanoid Robot Using Markerless Human Upper Body Pose Imitation. In *2019 IEEE-RAS 19th International Conference on Humanoid Robots (Humanoids).*

Yanlong Huang, Leonel Rozo, João Silvério, and Darwin G Caldwell. 2019. Kernelized movement primitives. *The International Journal of Robotics Research* 38, 7 (June 2019), 833–852. https://doi.org/10.1177/0278364919846363

Seyed Mohammad Khansari-Zadeh and Aude Billard. 2011. Learning stable nonlinear dynamical systems with gaussian mixture models. *IEEE Transactions on Robotics* 27, 5 (2011), 943–957.

Brigitte Krenn, Sepideh Sadeghi, Friedrich Neubarth, Stephanie Gross, Martin Trapp, and Matthias Scheutz. 2020. Models of Cross-Situational and Crossmodal Word Learning in Task-Oriented Scenarios. *IEEE Transactions on Cognitive and Developmental Systems* 12, 3 (2020), 658–668. https://doi.org/10.1109/TCDS.2020.2995045

Brigitte Krenn, Christiana Tsiourti, Friedrich Neubarth, Stephanie Gross, and Matthias Hirschmanner. 2019. Active Language Learning Inspired from Early Childhood Information Seeking Strategies. In *Workshop on Cognitive Architectures for Human-Robot Interaction at AAMAS 2019*.

Michael Lewis, Katia Sycara, and Phillip Walker. 2018. The role of trust in human-robot interaction. In *Foundations of trusted autonomy.* Springer, Cham, 135–159.

Timothy P. Lillicrap, Jonathan J. Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. 2016. Continuous control with deep reinforcement learning. In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings.*

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*. Springer, 740–755.

Matt MacMahon, Brian Stankiewicz, and Benjamin Kuipers. 2006. Walk the Talk: Connecting Language, Knowledge, and Action in Route Instructions. In *Proceedings of the 21st National Conference on Artificial Intelligence - Volume 2 (Boston, Massachusetts) (AAAI'06)*. AAAI Press, 1475–1482.

Ajay Mandlekar, Danfei Xu, Josiah Wong, Soroush Nasiriany, Chen Wang, Rohun Kulkarni, Li Fei-Fei, Silvio Savarese, Yuke Zhu, and Roberto Martín-Martín. 2021. What Matters in Learning from Offline Human Demonstrations for Robot Manipulation. In *5th Annual Conference on Robot Learning (CoRL)*. https://openreview.net/forum?id=JrsfBJtDFdI

Cynthia Matuszek. 2018. Grounded Language Learning: Where Robotics and NLP Meet.. In *Proceedings of the International Joint Conference on Artificial Intelligence.* 5687–5691.

Suraj Nair, Silvio Savarese, and Chelsea Finn. 2020. Goal-Aware Prediction: Learning to Model What Matters. In *Proceedings of the 37th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 119),* Hal Daumé III and Aarti Singh (Eds.). PMLR, Virtual, 7207–7219. http://proceedings.mlr.press/v119/nair20a.html

Takayuki Osa, Joni Pajarinen, Gerhard Neumann, J Andrew Bagnell, Pieter Abbeel, Jan Peters, and others. 2018. An algorithmic perspective on imitation learning. *Foundations and Trends in Robotics* 7, 1-2 (2018), 1–179. Publisher: Now publishers.

Guglielmo Papagni and Sabine Koeszegi. 2021. Understandable and trustworthy explainable robots: A sensemaking perspective. *Paladyn, Journal of Behavioral Robotics* 12, 1 (2021), 13–30. https://doi.org/doi:10.1515/pjbr-2021-0002

Alexandros Paraschos, Christian Daniel, Jan Peters, and Gerhard Neumann. 2018. Using probabilistic movement primitives in robotics. *Autonomous Robots* 42, 3 (March 2018), 529–551. https://doi.org/10.1007/s10514-017-9648-7

Affan Pervez, Yuecheng Mao, and Dongheui Lee. 2017. Learning deep movement primitives using convolutional neural networks. In *2017 IEEE-RAS 17th International Conference on Humanoid Robotics (Humanoids)*. IEEE, Birmingham, 191–197. http://ieeexplore.ieee.org/document/ 8246874/

Matthias Plappert, Marcin Andrychowicz, Alex Ray, Bob McGrew, Bowen Baker, Glenn Powell, Jonas Schneider, Josh Tobin, Maciek Chociej, Peter Welinder, et al. 2018. Multi-goal reinforcement learning: Challenging robotics environments and request for research. *arXiv preprint arXiv:1802.09464* (2018).

Pragathi Praveena, Guru Subramani, Bilge Mutlu, and Michael Gleicher. 2019. Characterizing Input Methods for Human-to-robot Demonstrations. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 344–353.

Astrid Rosenthal-Von Der Pütten and Nikolai Bock. 2018. Development and Validation of the Self-Efficacy in Human-Robot-Interaction Scale (SE-HRI). *J. Hum.-Robot Interact.* 7, 3 (Dec. 2018). Place: New York, NY, USA Publisher: Association for Computing Machinery.

Rouhollah Rahmatizadeh, Pooya Abolghasemi, Ladislau Boloni, and Sergey Levine. 2018. Vision-Based Multi-Task Manipulation for Inexpensive Robots Using End-to-End Learning from Demonstration. In *2018 IEEE International Conference on Robotics and Automation (ICRA).* IEEE, Brisbane, QLD, 3758–3765. https://doi.org/10.1109/ICRA.2018.8461076

Harish Ravichandar, Athanasios S. Polydoros, Sonia Chernova, and Aude Billard. 2020. Recent Advances in Robot Learning from Demonstration. *Annual Review of Control, Robotics, and Autonomous Systems* 3, 1 (May 2020), 297–330. https://doi.org/10.1146/annurev-control-100819-063206

Nicole L Robinson, Teah-Neal Hicks, Gavin Suddrey, and David J Kavanagh. 2020. The Robot Self-Efficacy Scale: Robot Self-Efficacy, Likability and Willingness to Interact Increases After a Robot- Delivered Tutorial. In *2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, Naples, Italy, 272–277. https://doi.org/10.1109/RO-MAN47096. 2020.9223535

Stéphane Ross, Geoffrey Gordon, and Drew Bagnell. 2011. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics.* 627–635.

Oliver Schürer, Benjamin Stangl, Christoph Hubatschke, and Christoph Müller. 2018. Experiments with a First Prototype of a Spatial Model of Cultural Meaning through Natural-Language Human-Robot Interaction. *Technologies* 6, 1 (Jan. 2018), 6. https://doi.org/10.3390/technologies6010006

Aran Sena and Matthew Howard. 2020. Quantifying teaching behavior in robot learning from demonstration. *The International Journal of Robotics Research* 39, 1 (2020), 54–72. Publisher: SAGE Publications Sage UK: London, England.

Pratyusha Sharma, Lekha Mohan, Lerrel Pinto, and Abhinav Gupta. 2018. Multiple interactions made easy (mime): Large scale demonstrations data for imitation. In *Conference on robot learning.* PMLR, 906–915.

Akira Taniguchi, Tadahiro Taniguchi, and Angelo Cangelosi. 2017. Cross-Situational Learning with Bayesian Generative Models for Multimodal Category and Word Learning in Robots. *Frontiers in Neurorobotics* 11 (Dec. 2017), 66. https://doi.org/10.3389/fnbot.2017.00066

Stefanie Tellex, Nakul Gopalan, Hadas Kress-Gazit, and Cynthia Matuszek. 2020. Robots That Use Language. *Annual Review of Control, Robotics, and Autonomous Systems* 3, 1 (May 2020), 25–55. https://doi.org/10.1146/annurev-control-101119-071628

Anna-Lisa Vollmer and Lars Schillingmann. 2018. On Studying Human Teaching Behavior with Robots: a Review. *Review of Philosophy and Psychology* 9, 4 (Dec. 2018), 863–903. https://doi.org/10.1007/s13164-017-0353-4

Sebastian Wallkötter, Silvia Tulli, Ginevra Castellano, Ana Paiva, and Mohamed Chetouani. 2021. Explainable Embodied Agents Through Social Cues: A Review. *J. Hum.-Robot Interact.* 10, 3, Article 27 (July 2021), 24 pages. https://doi.org/10.1145/3457188

David Whitney, Eric Rosen, Elizabeth Phillips, George Konidaris, and Stefanie Tellex. 2020. Comparing robot grasping teleoperation across desktop and virtual reality with ROS reality. In *Robotics Research*. Springer, 335–350.

Sarah Young, Dhiraj Gandhi, Shubham Tulsiani, Abhinav Gupta, Pieter Abbeel, and Lerrel Pinto. 2021. Visual Imitation Made Easy. In Proceedings of the 2020 Conference on Robot Learning (Proceedings of Machine Learning Research, Vol. 155), Jens Kober, Fabio Ramos, and Claire Tomlin (Eds.). PMLR, 1992–2005. https://proceedings.mlr.press/v155/young21a.html 6

Setareh Zafari, Isabel Schwaninger, Matthias Hirschmanner, Christina Schmidbauer, Astrid Weiss, and Sabine T. Koeszegi. 2019. ”You Are Doing so Great!” - The Effect of a Robot’s Interaction Style on Self-Efficacy in HRI. In *2019 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE.

Stefan Zahlner, Matthias Hirschmanner, Timothy Patten, and Markus Vincze. 2020. Teleoperation System for Teaching Dexterous Manipulation. In *Workshop on Hand-Object Interaction: From human demonstrations to robot manipulation at RO-MAN2020.*

Tianhao Zhang, Zoe McCarthy, Owen Jow, Dennis Lee, Ken Goldberg, and Pieter Abbeel. 2018. Deep imitation learning for complex manipulation tasks from virtual reality teleoperation. In *IEEE International Conference on Robotics and Automation (ICRA).*

Trustworthy Robots in Work Context

Bottom-Up Research on Assistive Robots for the Aging Population

Isabel Schwaninger, Astrid Weiss, Geraldine Fitzpatrick

Abstract

Robots are developed in the hope to solve various problems that our societies currently face. One of the most pressing problems is the aging population, placing increasing pressure on the care system as people live longer, and often with chronic diseases. Robots are also considered as a possible solution to this problem. However, imaginations of the role of robots are frequently driven by technology-utopian top-down agendas without regard to practical realities of everyday life of the older adults and other stakeholders they seek to support. This chapter presents an overview of assistive technology for the aging population, along with building blocks for Human-Robot Interaction (HRI) research in these contexts, and the challenges that arise. On this basis, we characterize ways of conducting bottom-up research to explore trustworthy HRI for older adults in home environments.

Keywords

human-robot interaction, aging, home, bottom-up research, aging in place, trust

1 Introduction

Robots and artificial intelligence (AI) are being developed to solve various problems that our societies currently face. One of these problems is associated with the aging population. The demographic changes placing pressure on care systems [World Health Organization 2020b], as people live longer and often with chronic diseases. In order to support people's independent daily living and to also meet the increasing demand of caregivers and for aging in place, there is a trend to envision assistive robots as a next generation of technical solutions for Active and Assisted Living (AAL). Studies have already demonstrated positive effects of robots on people's health and well-being in living spaces [de Graaf et al. 2015; Klamer and Allouch 2010; Tsiourti et al. 2014; Wada et al. 2005, 2004; Broadbent et al. 2016].

However, despite technical advancements and many years of research on a wide range of AAL technologies since the 1990s [Haslwanter et al. 2020] – including safety systems, security, monitoring, communication, and entertainment systems, and home automation [Turner and McGee-Lennon 2013; Haslwanter et al. 2020]) –, there is still a limited uptake even of basic AAL technology on the market [Haslwanter and Fitzpatrick 2017], let alone more complex solutions such as assistive robots.

One reason for this limited uptake of assistive technologies might be related to the issue of trust. Previous research suggests that older adults use a language indicating distrust in relation to digitalization [Knowles and Hanson 2018]. Similarly, trust is also considered a critical topic in Human-Robot Interaction (HRI) research

https://doi.org/10.34727/2022/isbn.978-3-85448-052-5_9

[Billings et al. 2012; Schaefer 2016; Mcknight et al. 2011]. A lack of trust has also been identified as a barrier for older adults using health information technologies [Fischer et al. 2014], which can therefore be anticipated as a challenge for assistive robots in this context.

Another reason for an anticipated limited uptake of assistive robots might be related to unrealistic conceptualizations of human-robot constellations and older people at home. Previous research has revealed discrepancies between what older adults expect from future assistive robots and what state-of-the-art technology can actually deliver [Weiss and Spiel 2021; Vincze et al. 2014]. Given that AAL research has been going on since the 1990s, there is a growing body of research applying and studying these technologies in use also outside of controlled laboratory settings [Lussier et al. 2020; Turner and McGee-Lennon 2013; Masterson Creber et al. 2016; Axelrod et al. 2009]. From this research, we are gaining a growing understanding of the contextual factors regarding putting these technologies to work. Nevertheless, to date, studies on assistive robots in relation to trust have largely been lab-based [Schwaninger et al. 2019] and studies outside of the lab are often too short and/or focus only on specific aspects, such as usability and technology acceptance (e.g., [Bajones et al. 2019]), but not addressing the whole adoption process. Therefore, we do not have any concrete insights on the role of trust and the impact of contextual factors for assistive robots in older people's homes, yet.

This chapter presents a review of recent related work on assistive technology and robots (being a recent instance of assistive technology) to support aging at home. First, we present a section on assistive technology for the aging population, discussing related work on tech- nology for active and assisted living, robots as a recent instance of these, and perspec- tives on aging. We then discuss building blocks for bottom-up research on assistive robots. These building blocks include related work on how to understand living spaces, and people and stakeholders there. We follow with a section where we propose to explore trustworthy HRI for older adults in home environments taking bottom-up approaches. Here, we also describe various case studies (to be found in referenced publications) on investigating peo- ple's practices, on co-creation methods for robots, and on longitudinal studies with robots. We close the chapter with a Conclusion section.

2 Assistive Technology for the Aging Population

AAL technologies have been promoted for many years for aging in place. In this section, we aim to gain a better understanding of assistive technology for the aging population. We therefore discuss related work on technology for active and

assisted living, on robots as a recent instance of assistive technology, and on perspectives on aging as such.

2.1. Technology for Active and Assisted Living

AAL technologies have been promoted for many years [Haslwanter et al. 2020; Choukou et al. 2021] as a way to meet the desire among older adults to stay healthy and live autonomously in their homes for as long as possible [Bieg et al. 2022; Peek et al. 2015; Liu et al. 2016; Bloom and Luca 2016]. Their goals are among others, to enable an independent, active, and self-determined life [Vimarlund et al. 2021; Brauner and Ziefle 2021; Nilsson et al. 2021; Dupuy et al. 2016], stay socially connected [Schomakers et al. 2018; Blackman et al. 2016] or to feel safer in everyday life at home [Turjamaa et al. 2019]. Relevant technologies include safety systems, systems for security, monitoring the health status [Lussier et al. 2020], communication, and entertainment, as well as home automation [Turner and McGee-Lennon 2013; Haslwanter et al. 2020]). Notably, that the term AAL has also been motivated by a research funding scheme of the European Union for *Aging Well in the Digital World*[1]. Furthermore, the term – given its broad scope – overlaps with related terms such as smart home technologies or gerontechnologies, and the distinction between these terms is not clear-cut.

AAL technologies are either implemented as individual services or comprehensive systems that combine a number of different services (multi-service systems). While specific contexts of application, use cases, and desired outcomes regarding AAL technologies are highly diverse, they share two general characteristics: First, they are ambient, meaning that they are seamlessly integrated into people's environment, realized through a wide array of different embedded technologies [Schomakers et al. 2018], such as camera and sensor systems integrated into the immediate home environment, wearable devices [Correia et al. 2021], or smart everyday-objects [Cicirelli et al. 2021]. Second, they assist people. Examples include the implementation of a voice-controlled smart home environment, designed to support people with visual impairments [Vacher et al. 2015] (where smart homes can also involve robots [Do et al. 2018]; more about these later), or rehabilitation technologies designed to assist people by motivating them and promoting exercises after a stroke [Axelrod et al. 2009].

Blackman et al. [Blackman et al. 2016] identified three generations of AAL. The first generation of AAL technologies includes community, social, and personal response systems, mostly designed as wearable devices that can be used to trigger an alarm to contact a person in a 24-h call center. The benefit is potentially decreased stress levels among older adults and their family and caregivers;

1 https://www.aal-europe.eu/

as well as the ability to live at home longer. Meanwhile, a disadvantage is that the person actually must wear the device, which can be stigmatizing, and also remember to do so in high-risk situations such as when getting up at night. The second generation of AAL technologies is characterized by integrated electronic components which not only respond to, but also detect emergencies with sensors [Blackman et al. 2016], such as a fall, or environmental hazards [Sixsmith 2000]. These technologies are used within the home only, and they may feel intrusive. The third generation of AAL technologies combines the benefits of earlier technologies, aiming to detect and report problems and also prevent them. By integrating computing systems and assistive devices, such as wearable and environmental sensors into living spaces, they monitor both the environment and the older person. A potential benefit is also reduced stigma associated with monitoring and assistance by embedding technology within everyday objects and hiding them [Blackman et al. 2016].

2.2. Robots as a recent Instance of Assistive Technology

A recent instance of AAL are assistive robots, which are embodied agents[2]. The variety of tasks that such assistive robots are envisioned to take over is manifold. In the broader scope of healthcare, one of the application areas includes medical robotics. Medical robots are being used increasingly, for example, to support surgical procedures [Nwosu et al. 2019]. There are robots used for pain relief [Azeta et al. 2018], and studies propose assistive uses of robots in dementia or care of older adults [Nwosu et al. 2019]. Robots are also proposed to enhance health and psychological factors of people in general and older adults in particular, for example, by providing companionship [Cifuentes et al. 2020]. Researchers have worked on robotics for tele-healthcare [Azeta et al. 2018], for instance providing assistance and being remotely operated by a doctor [Martinez-Martin and del Pobil 2017]. Health-related applications also include robots for rehabilitation (e.g. Auto Ambulator), including neuro-rehabilitation [Krebs et al. 2021]. Other work has proposed mobile robots to aid physiotherapists in their work [Gerling et al. 2016]. There are also various applications of AI in healthcare, for example, processing and analyzing patient data [Amisha et al. 2019]; and while these do not necessarily require robots per se, they may assist doctors in primary patient care.

Other types of service robots were also proposed for home environments. They can be very broadly defined as “assistive devices designed to support people living independently” [Martinez-Martin and del Pobil 2017], for example, by

2 Note, what constitutes a robot is not clear cut, especially in comparison to other assistive technology. Robots are often characterized as embodied agents [Feil-Seifer and Matarić 2009]; and they are also treated as a separate entity in academic disciplines, see e.g. https://humanrobotinteraction.org

assisting with mobility, household tasks, and monitoring safety and health [Martinez-Martin and del Pobil 2017]. Because these robots need to adapt to the living conditions to some extent, they require a certain degree of complexity [Martinez-Martin and del Pobil 2017]. Furthermore, they are embodied. They can assist in mobility (such as Friend II), or support in fetching and carrying (such as Boltr). Robots have been designed for personal care, (e.g. Bestic) and for cleaning (e.g. the vacuum cleaner robot Scoooba) [Werner et al. 2015]. They can be also intended for older adults to feel safer and stay longer in their homes by providing fall prevention measures, as well as emergency detection and handling [Martinez-Martin and del Pobil 2017; Bajones et al. 2018].

Assistive robots can offer functionalities for social purposes, for example, telepresence robots to connect to other people (e.g. Giraff). Companion robots, such as Hector and the seal robot Paro, are intentionally designed as emotional agents [Werner et al. 2015]. They are designed to proactively assist older adults in everyday tasks, reduce stress and promote well-being, to enhance social interaction and elicit emotional responses [Martinez-Martin and del Pobil 2017]. Potentially, companion robots also include entertainment robots (e.g. Ifbot) [Werner et al. 2015], or social robots for therapy and care [Cifuentes et al. 2020]. As an example of social robots, pet-like robots are proposed to increase well-being of patients during hospital stays [Cifuentes et al. 2020]. Similarly, baby-type robots are designed for being taken care of an older person requiring nursing care, as part of Babyloid-centered therapies for promoting motivation to older adults [Martinez-Martin and del Pobil 2017].

2.3. Perspectives on Aging

While assistive technologies aim to target older adults, there is no universal agreement on what aging and age (in particular old age) actually mean. There are furthermore various perspectives across cultures and generations [Palmore 1999], and across research disciplines. Broadly, the different approaches and understandings of aging can be found under the umbrella term ‘gerontology’. Different conceptualizations of aging can be either implicitly or explicitly embodied into technology design [Harley 2011; Fitzpatrick et al. 2015], which is why we aim to reflect on the forming of these conceptualizations.

Different definitions of aging are used at the policy level. The World Health Organization (WHO) promotes Active Aging [World Health Organization 2002], where aging is regarded as the process of optimizing opportunities for health, participation, and security in order to enhance the quality of life over time [World Health Organization 2018; Foster and Walker 2015]. Active Aging applies to individuals and groups, as well as entire populations [World Health Organization

2018]. Another term employed by the WHO is Successful Aging [Bowling and Dieppe 2005], which also has a proactive, but rather normative approach. It consists of three elements, including the reduction of disease and disability, maintenance of high cognitive and physical functioning, and active engagement with life [Rowe and Kahn 1997]. A term that is being used also by the WHO more recently is Healthy Aging, which places focus on "creating the environments and opportunities that enable people to be and do what they value throughout their lives" [World Health Organization 2020a]. Not having any physical or mental diseases is not a requirement for healthy aging, as it is more about how these conditions are handled to support well-being, to enable older adults to remain a resource (e.g., to their families, communities and economies); with a particular focus on creating environments that minimize the exposure to health risks, access to quality health and social care, and a focus on the opportunities brought by aging [World Health Organization 2020a].

While these perspectives on aging at a policy level seek to promote a rather active lifestyle, participation and quality of life, other perspectives on aging have a different focus. A dominant view tends to conceptualize aging from a bio-medical or from a social point of view. Here, aging is also associated with an accumulation of loss, and "an ongoing 'diminishment' of function" [Fitzpatrick et al. 2015]. From a bio-medical point of view, aging is reflected in physical, biological and cognitive aspects of functioning, and it is also addressed through various types of healthcare services [Harley 2011]. As a person's cognitive performance changes over time, the neurobiological changes are regarded as a decline, for instance, due to losses of cognitive capacity, like memory. Besides these bio-medical models, there are social models of aging: Activity theory is intrinsically linked to a loss of participation in society and in social roles beyond retirement [Harley 2011; Fitzpatrick et al. 2015]. From this perspective, aging adults who engage in daily activities that they perceive productive age successfully. They may, for example, engage in volunteering, care-giving and self-development [Karim et al. 2018]; activities in which there is furthermore a value of social interaction. According to this perspective, people's level of activity is further linked to life satisfaction, which also affects a person's view on themselves (self-concept) [Diggs 2008]. Another social model of aging is disengagement theory. From this point of view, disengagement is regarded as an adaptive response to aging, and increasing social withdrawal with age is regarded as normal and healthy. Older adults voluntarily transfer the power to younger generations; which is even regarded as beneficial for both the aging society and individuals [Diggs 2008]. Both bio-medical models and social models of aging place emphasis primarily on the deficits of aging [Harley 2011; Fitzpatrick et al. 2015], and they see older people as rather passive recipients of social and medical intervention. They ignore the subjective experience and individual adaptations made in day-to-day life.

In contrast to a focus on decline, aging as adaptation is covered in other studies. Here, aging is regarded as a positive developmental lifespan process [Erikson and Erikson 1998]. The model of Lifespan Development postulates eight successive stages of individual human development that are influenced by biological, psychological and social factors throughout the lifespan [Orenstein and Lewis 2020]. According to Erkison's Lifespan development, a person's identity is adapted in line with one's life stage. Middle and late adulthood in particular are regarded as relevant, because active and significant personality development also takes place at these stages. Further, the stage of old age is concerned with a conflict between integrity and despair or disgust, where the individual looks back and reflects, also gaining wisdom [Erikson and Erikson 1998]. Tornstam [Tornstam 2005] extended Erikson's lifespan development with an additional stage in life, Gerotranscendence. In this additional stage, individuals tend to become less self-occupied, increasingly feeling affinity with past generations, and decreasingly interested in superfluous social interaction. They may also experience a decreased interest in material things, and positive solitude becomes increasingly important to them [Tornstam 2005].

Another perspective suggests that developmental opportunities of "successful aging" take place when there is a compensation for age-related declines by developing other capacities, namely by selectivity with optimization and compensation [Baltes and Baltes 1990]. From this point of view, older adults can promote their quality of life by selecting particular life goals over others. They can acquire and coordinate personal resources for selected goals (optimization) and employ alternative means to reach a certain goal (compensation) [Harley 2011]. According to Socio-emotional Selectivity, perceived proximity of death can affect older adults' selectivity. The perspective is based on the assumption that social contact is motivated by a range of different goals, ranging from basic survival to psychological goals. The importance of these goals fluctuates depending on our age, in particular, emotional regulation increases with older age, while the acquisition of information, and the desire to affiliate with unfamiliar people becomes less important. Therefore, socio-emotional selectivity triggers increasing emotionally meaningful and socially-oriented goals. In this manner, older adults avoid superficial social contact and seek to deepen intimacy [Carstensen 1992; Carstensen et al. 1999]. Joyce and Loe argued that older adults are active adaptive agents to technology. They argue that this group of people does not consist of passive consumers, but "technogenarians" who creatively utilize and adapt technological artifacts to fit their needs [Joyce and Loe 2010]. With this in mind, assistive technology must be designed in a way that people are able to adapt it to their specific needs, also taking into account their home environment and stakeholders. Specifying these needs requires certain building blocks to be considered in the design process, as outlined below.

3 Building Blocks for Bottom-Up Research on Assistive Robots

While assistive technology has been developed since many years for aging in place, there are certain building blocks that need to be considered apart of the technology itself for aging. Given that bottom-up research as it is presented here involves an engagement with situated experiences, such building blocks include at least living spaces (such as different types of homes), and people and stakeholders living and working in these environments. These two building blocks are characterized in the following.

3.1. Understanding Living Spaces

Different types of living spaces are often implicitly or explicitly considered for aging, such as private homes and institutional care homes, where home is certainly a culture-specific term. According to previous research, home can be regarded as an abstract concept related to a wide set of associations and meanings. It is a physical space with subjective attachments to it [Pani-Harreman et al. 2021], which holds a symbolic meaning [Moore 2000]. The multifaceted aspects of home can be described as “a place, a relationship and an experience” [Gillsjö et al. 2011]. In a study conducted with older adults in particular, home has been conceptualized as a place that has been built together for a long period of time, a relational place, a place “closest to the heart” [Dahlin-Ivanoff et al. 2007], an experience, and freedom. It has been associated with security (due to the familiar neighborhood, memories and functionalities), and freedom (being a place for reflection, a social meeting-point, and leaving your own mark) [Dahlin-Ivanoff et al. 2007]. It is associated with belonging, and the experience of home of older people in particular has been associated with a movement between the well-known present and the unknown future (i.e., as there may be a day where one has to leave home) [Gillsjö et al. 2011]. Moore also points out that the following basic terms have been frequently associated with home: privacy, security, family, intimacy, comfort, and control [Moore 2000]. Studies concerning changes of home due to relocation, aging or physical or/and cognitive frailty [Leith 2006; Renaut et al. 2015; Case 1996] indicate that aging in a familiar environment is also likely to have a positive impact on the wellbeing of older adults in later life [Pani-Harreman et al. 2021; Van Dijk et al. 2015]. Certainly, aging at home is also connected to financial aspects like the opportunity to receive care at a lower cost [Pani-Harreman et al. 2021].

Conceptualizing home is important, because of the way in which technology is envisioned to be used at home, which may have an effect on design choices. In-

novations may have changed the way we perceive home, as well as the physical quality of the home itself. For example, telephones used to be an important part of home, often situated in an easily accessible place. This communication spot at home has now become more flexible, through the use of mobile devices. When considering technology installed for telehealth, research suggests that rehabilitation technologies can have an impact both on physical arrangements of the home and on how home is perceived and felt, which must be considered when designing these technologies [Axelrod et al. 2009]. With an increasing number of digital applications to be used at home, the quality of the home as a place can change. With an increasing use of health-related applications, home can be perceived as an extended care facility [Boyne and Vrijhoef 2013]. A potential risk is designing technical artifacts proscribing fragile, home-bound users, where older adults are envisioned to be bound to their physical homes through the use of assistive technology. In contrast, people may want to maintain their social networks also in places outside the home [Aceros et al. 2015]. If robots are designed for home environments, home as a place and associated home practices need to be taken into account to promote quality of life. For example, home organization is relevant to consider for HRI [Cha et al. 2015], and as are power relations [Lee et al. 2017a]. These are also connected to people and stakeholders at home, which are further characterized in the following

3.2. Understanding People and Stakeholders

Another issue to consider is the people who live and work in home environments. AAL technologies are intended to support older adults while aging in place [Choukou et al. 2021]. The group of older adults is, however, quite diverse in itself, and people within the same age cohort may have drastically different needs. Therefore, technological solutions aim to either target a broad spectrum of people, resulting in rather complex systems with a high degree of functionalities, or people with very specific needs (e.g., to support or promote physical or psycho-social health). The target group is also often referred to as primary users [Werner et al. 2015]; the term “user” has been debated in previous research [Bannon 1995].

Besides the group of older adults, other people are involved in the interaction with assistive technologies. As mentioned above in the case of telemonitoring for patients with chronic heart failure [Boyne and Vrijhoef 2013], including in the home environment, older adults are in contact with other people even if they live on their own. These – often called secondary users [Werner et al. 2015] – include all kinds of peers, extended family, or care workers, such as informal or formal care workers, sometimes also referred to as caregivers. There exists work on proposing robots to support secondary users, such as formal care workers

in institutional settings [Johansson-Pajala et al. 2020], or informal care workers (where some work also involved long-term co-design activities [Moharana et al. 2019]). Some work has aimed to design trustworthy care robots [Stuck and Rogers 2018], where such work has emphasized the direct role of robots as care workers (interestingly, rather than the robots' potential for supporting the work practices of human care workers, for example).

The caregivers' profession is relevant for envisioning robots in home environments, especially when it comes to institutional care homes. Recently, workers have been confronted with a growing number of challenges, including low recognition of one's contribution, inadequate pay and workload, strong emotional experiences and increasing work-related stress and burnout [Foà et al. 2020]. In contrast, working autonomy, professional growth, positive relationships with colleagues and older adults increase job satisfaction [Foà et al. 2020]. Many care homes are understaffed, which forces caregivers in Europe to work overtime [Foà et al. 2020], as is the situation for hospital nurses, with a reported decrease in patient safety and quality of care, or even care left undone [Griffiths et al. 2014]. Missed care in the medical context again not only leads to decreased patient satisfaction, but it can also lead to medical problems like medication errors, urinary tract infections, patient falls, pressure ulcers, care quality and patient readmissions [Recio-Saucedo et al. 2018]. A response to these professional burdens is to support care workers in providing healthcare assistance, in performing daily tasks, or in the increase of self-management [Martinez-Martin and del Pobil 2017]. Despite the evident role of care workers, especially in institutional home settings, their role has been rarely considered in action. Few exceptions include Hornecker et al. who have referred to the triadic interaction between a robot, a care worker, and an older person, and hence constitutes one of the very few multi-actor approaches [Hornecker et al. 2020]. They argued that such interaction can only be satisfyingly designed when all parties are taken into account.

Similarly, tertiary users have been hardly addressed in relation to assistive robots [Werner et al. 2015]. These include service providers, installation and maintenance technicians, insurance companies, municipalities, architects, social agencies, and guarantors of privacy, safety, and ethical procedures [Johnson et al. 2014]. Certainly, their needs and preferences can be very different from the needs of primary or secondary users (e.g., given the financial aspects that are also sometimes related to receiving care [Pani-Harreman et al. 2021]).

4 Exploring trustworthy HRI for Older Adults in Home Environments

As discussed above, a variety of assistive technologies has been developed since the 1990s for aging at home, including robots as a recent instance. Furthermore, different types of living spaces that involve different people/stakeholders must be considered; and different perspectives on home and aging that are reflected in technology design may open novel design spaces. For example, a robot could prescribe a home-bound person, or it could promote a more active lifestyle and choices, also becoming part of an environment that promotes community or autonomy.

Despite these various perspectives entertained in previous studies, challenges remain in practice. We described only a few of these challenges in Section 1, including a lack of marketable AAL products despite numerous years of research, as well as trust as a critical topic related to the limited uptake.

The integration of assistive robots into everyday life depends heavily on design choices. While some of these choices are technical or directly related to the design of a technical artifact, such as when defining how a robot should respond to contextual factors [Rosenthal-von der Pütten et al. 2020], others may be not be limited to technology itself. However, when developing robots in a top-down manner, many of these opportunities for exploring different design approaches are potentially being missed. Furthermore, challenges in integrating robots in real-world settings are unlikely to be revealed in laboratory HRI studies or in observations of short-term interactions even when using authentic settings [de Graaf et al. 2015]. For example, while technical readiness is a key requirement for robots, it is also necessary to understand processes in the real world that robots are intended to assist with, for example, the manner in which people live and work. Understanding people's values is likewise important, such as what autonomy means for people [Hornung et al. 2016], or the specific issues that people raise in relation to trust or distrust. If these issues are only fully understood in practice, after the robot has already been built, then the opportunities for change or re-design are limited or very costly. A bottom-up approach engenders an earlier engagement with people and their context when designing technology [Broadbent et al. 2016], and this can potentially clarify these sorts of problems earlier to save time and costs later. Here, different approaches have been proposed, such as placing focus on people's social practices [Wulf 2009; Wulf et al. 2011; Kuutti and Bannon 2014; Ganglbauer et al. 2013; Schmidt 2018], participatory approaches [Lee et al. 2017b; Lan Hing Ting et al. 2018; Frennert et al. 2012; Weiss and Spiel 2021], or long-term studies with robots [Irfan et al. 2019, 2021; de Graaf et al. 2017, 2018]

In the following, we propose different ways of understanding and designing assistive robots (and trustworthy HRI) by engaging with the context of older adults. To understand current issues bottom-up and to avoid re-inventing the wheel, we propose to learn from people's current practices and their use of current assistive technologies to understand the challenges and to provide lessons learned for designing robots for this context. Furthermore, we propose methodological explorations for participatory design to explore ways of conceptualizing robots in people's living spaces in relation to their social practices. Here, a closer look at trust is promising, especially in home environments and in relation to home practices. Further, building on existing work, we propose long-term studies to be conducted with robots to support people working in living spaces.

4.1. Investigating People's Practices

One approach to conduct bottom-up research with robots is a focus on people's practices. In Human-Computer Interaction (HCI), a shift from interactional research to practice-based research in the everyday is also reflected in the turn to practice [Kuutti and Bannon 2014]. While early methods in HCI were inspired by psychological sciences involving controlled short-term, lab-oriented studies, which are according to Kuutti & Bannon [Kuutti and Bannon 2014] embedded in the Interaction paradigm, this is not the case in the recently emerging Practice paradigm. In previous practice-oriented work, the practical accomplishment and "dynamic and situated 'interactional' aspects [...] to be accounted" [Fitzpatrick 2003, p. 91] was highlighted. Generally, practice approaches explore "[...] historical process and performances, longer-term actions which persist over time, and which must be studied along the full length of their temporal trajectory[,][...] situated in time and space"[Kuutti and Bannon 2014, p. 3543]. Further, the broader context is taken into account, and it is "intervowen within the practice" [Kuutti and Bannon 2014, p. 3543].

As a starting point, there is an opportunity to choose a topic like trust, a topic which is known to be important in HRI [Billings et al. 2012] and in the context of older adults and technology [Knowles and Hanson 2018; Fischer et al. 2014], and to investigate it across the literature. There may be other research areas that have a longer tradition for understanding people's practices [Kuutti and Bannon 2014], which is an opportunity for mutual learning. Research in the related area of Computer Supported Cooperative Work (CSCW) has taken a practice approach in collaborative care engagements [Fitzpatrick and Ellingsen 2013]. While CSCW research does not necessarily focus on robots, we took a critical look at the literature to identify research gaps in HRI and CSCW specifically in relation to trust

and robots, investigating what CSCW can can contribute to an understanding of trust in the field of HRI [Schwaninger et al. 2019].

Moving from the literature to the real world, there are also opportunities to investigate people's use of AAL technologies (e.g., sensor-based technologies or technology for communication) to provide lessons learned for designing robots. This can be done in private households, as we did in other work [Schwaninger et al. 2020]. Taking a socio-technical perspective, we mapped out relationships between older adults and various actors in a network, pointing to relatedness needs and how technologies were integrated in relationships [Schwaninger et al. 2020]. Furthermore, since the COVID-19 pandemic, there has been also an increase in technology usage in various areas other than care [Marston et al. 2020; Cmentowski and Krüger 2020; Mastrianni et al. 2021], where previous research has also emphasized the potential of technology as a response to isolation in the care context that has occurred [Gallistl et al. 2021]. While COVID-related restrictions have been ordered top-down by governments, earlier studies have shown how healthy aging is strongly perceived as an active achievement that includes the management of lifestyles, health and illness, and the active balancing of social life and financial and material circumstances [Sixsmith et al. 2014]. Due to these tensions, we used the opportunity to explore how the pandemic has triggered the usage of technology and the readiness to engage with current or future AAL technologies in different types of homes, and further how it has affected aging and associated experiences. We found that there was an impetus to enhance digital literacy triggered by COVID experiences, and increasing workload associated with new ways of putting technology to care work [Schwaninger et al. 2022]. There is also a potential for the use of robots that are similar to communication technology like tablets, which have so far more been accessible in everyday life compared to robots in living spaces. However, the work that is required to make use of robots as part of care work certainly needs to be tackled.

4.2. Exploring Co-creation Methods for Robots with Older Adults

Given assistive technology needs to fit into complex realities of older adults, participatory design has been increasingly promoted and recognized as an "important route to context-sensitive, person-centered and sustainable health innovation" [Langley et al. 2019, p. 3] for older adults. Recent HRI research has also explored participatory methods for designing robots for older people and care contexts [Frennert et al. 2013; Lee et al. 2017b; Lan Hing Ting et al. 2018; Georgiou et al. 2020; Rogers et al. 2021]. Participatory design can support designers in developing robots that meet older adults' needs, capabilities and preferences

on the one hand [Rogers et al. 2021], and promote mutual learning between researchers and participants [Lee et al. 2017b] on the other.

While participatory design has been conducted in HCI for some time, in HRI, it is still relatively new and comes with specific challenges [Weiss and Spiel 2021]. As robots are technically complex, the involvement of older adults, for example, in building prototypes is not straight-forward. While co-designing screens, for example, could be done with pen and paper, building a robot prototype requires technical skills. Therefore, the co-design process itself also involves multiple people and, consequently, their perspectives [Rogers et al. 2021]. In recent participatory design studies, Lan Hing Ting et al. [Lan Hing Ting et al. 2018] used ethnographic methods to explore the co-design and evaluation process of a mobile social robotic solution for older adults following a living lab approach, involving the people considered to be primary users, sociologists, designers, and engineers. Furthermore, the use of robot prototypes can be beneficial for co-design to involve older adults with actual systems that they can discover [Lee et al. 2017b] and potentially extend. In prototyping workshops however, Bråthen et al. [Bråthen et al. 2019] found that developing a story about a robot in the context of older people's homes and in the daily life of older adults is essential for successful design and prototyping.

A challenge in participatory design or co-design for robots with older adults is that several stakeholders are involved, often working as interdisciplinary teams [Lan Hing Ting et al. 2018] (e.g., gerontologists, social scientists, engineers). To make collaboration in these teams more effective, there are opportunities to support such teams with tools [Axelsson et al. 2021]. Therefore, we started to develop a method to facilitate a shared understanding of older adults' everyday life in ideation phases. Because trust is a critical topic in HRI, and as older adults raised privacy concerns in other studies [Schwaninger et al. 2020], trust can be used as an icebreaker to facilitate conversations on opportunities for assistive technology to support older adults at home, along with other contextual elements and playful activities [Schwaninger et al. 2021].

4.3. Conducting long-term Studies with Robots

While learning from current technologies for robots and developing methods for cocreation with older adults are useful building blocks for bottom-up research, they do not involve any long-term experiences with actual robots. This can have advantages, e.g, in design [Schwaninger et al. 2021], as it allows ideation with fewer pre-assumptions of what actually constitutes a robot. It can further be useful to focus on needs rather than technical readiness, especially in co-creation processes. Nevertheless, studies with actual robots are also essential in the de-

sign and development process, revealing the opportunities that come with robots (i.e., the potential for different functionalities as described earlier), and providing new opportunities for supporting people. Further, technology, however it is designed, could also change people's practices[3], and it is therefore important to investigate the use or non-use of robots in real life. Studies have also shown that the novelty effect can decrease after some time of usage, as people get used to a robot and as it can become repetitive and predictable [Portugal et al. 2019; Martinez-Martin and del Pobil 2017]; and other work has provided valuable insights on the non-use of robots in homes and associated findings on motivation, as well as user types and needs [de Graaf et al. 2017]. The opportunity to conduct long-term studies with robots is an important building block of bottom-up research (i.e., to look beyond the use of off-the-shelf technologies).

Previous studies that have taken a long-term approach for aging at home either focus on older adults living in private homes [Bajones et al. 2018; de Graaf et al. 2017] or on the residents living in care homes. For example, de Graaf et al. [de Graaf et al. 2015] have provided insights on people's attitudes and relationship-building with or toward robots in private homes. Several studies that have taken place in institutional care homes suggest that care workers facilitate the interaction between robots and residents. Further, robots can enhance the quality of care, providing care workers with an additional tool to work with residents [Carros et al. 2022].These and other studies are needed to gather insights about long-term effects on the work that is needed to actually put a robot to work in living spaces.

5 Conclusion

Several challenges for human-centered research of assistive technology in the context of the aging population in general and assistive robots specifically, must be addressed bottom-up rather than top-down. We therefore presented an overview of related work on assistive technology for the aging population, along with building blocks that are further relevant to understand some of the current challenges bottom-up. Following this discussion of related work as a relevant basis, a variety of qualitative and explorative approaches have been briefly proposed based on our work [Schwaninger et al. 2020, 2021; Carros et al. 2022; Schwaninger et al. 2022], contributing to better understand people's needs and impact factors of actual technology uptake. The characterization of our work includes the investigation of people's practices [Schwaninger et al. 2019, 2020,

3 It should be noted at this point that practice approaches as described above can also involve long- term studies with robots, as these studies provide opportunities to observe, change, understand and design for people's practices.

2021], the exploration of co-creation methods [Schwaninger et al. 2021], and long-term studies with robots after deployment in their intended context-of-use [Carros et al. 2022]. Detailed insights from these studies are provided in the related publications and in the PhD thesis of Isabel Schwaninger (forthcoming in 2022). By exploring these bottom-up approaches, we aim for an understanding and design of HRI with assistive technologies that are considered as trustworthy by various stakeholders.

Bibliography

Juan C Aceros, Jeannette Pols, and Miquel Domènech. 2015. Where is grandma? Home telecare, good aging and the domestication of later life. *Technol. Forecasting Social Change* 93 (Apr 2015), 102–111. https://doi.org/10.1016/j.techfore.2014.01.016

Amisha, Paras Malik, Monika Pathania, and Vyas Kumar Rathaur. 2019. Overview of artificial intelli- gence in medicine. *J. Family Med. Prim. Care* 8, 7 (Jul 2019), 2328. https://doi.org/10.4103/ jfmpc.jfmpc_440_19

Lesley Axelrod, Geraldine Fitzpatrick, Jane Burridge, Sue Mawson, Penny Smith, Tom Rodden, and Ian Ricketts. 2009. The reality of homes fit for heroes: design challenges for rehabilitation tech- nology at home. *Journal of Assistive Technologies* (Jul 2009). https://doi.org/10.1108/ 17549450200900014

Minja Axelsson, Raquel Oliveira, Mattia Racca, and Ville Kyrki. 2021. Social Robot Co-Design Canvases: A Participatory Design Framework. *J. Hum.-Robot. Interact.* 11, 1 (2021), 1–39. https://doi.org/10.1145/3472225

Joseph Azeta, Christian Bolu, Abiodun A Abioye, and Oyawale Festus. 2018. A review on humanoid robotics in healthcare. *MATEC Web of Conferences* 153, 5 (Jan 2018), 02004. https://doi.org/10.1051/matecconf/201815302004

Markus Bajones, David Fischinger, Astrid Weiss, Puente Paloma De La, Daniel Wolf, Markus Vincze, Tobias Körtner, Markus Weninger, Konstantinos Papoutsakis, Damien Michel, Ammar Qammaz, Paschalis Panteleris, Michalis Foukarakis, Ilia Adami, Danae Ioannidi, Asterios Leonidis, Margherita Antona, Antonis Argyros, Peter Mayer, Paul Panek, Håkan Eftring, and Susanne Frennert. 2019. Results of Field Trials with a Mobile Service Robot for Older Adults in 16 Private Households. *ACM THRI* (Dec 2019). https://dl.acm.org/doi/10.1145/3368554

Markus Bajones, David Fischinger, Astrid Weiss, Daniel Wolf, and Susanne Frennert. 2018. Hobbit: Providing Fall Detection and Prevention for the Elderly in the Real World. *Journal of Robotics* 2018 (Jun 2018), 1–20. https://doi.org/10.1155/2018/1754657

Paul B Baltes and Margret M Baltes. 1990. Psychological perspectives on successful aging: The model of selective optimization with compensation. In *Successful Aging: Perspectives from the Behavioral Sciences.* Cambridge University Press, Cambridge, England, UK, 1–34. https://doi.org/10.1017/CBO9780511665684.003

Liam J Bannon. 1995. From Human Factors to Human Actors: The Role of Psychology and Human-Computer Interaction Studies in System Design. In *Readings in Human–Computer Interaction.* Morgan Kaufmann, 205–214. https://doi.org/10.1016/B978-0-08-051574-8.50024-8

Till Bieg, Cornelia Gerdenitsch, Isabel Schwaninger, Bettina Manuela Johanna Kern, and Christopher Frauenberger. 2022. Evaluating Active and Assisted Living technologies: Critical methodological reflections based on a longitudinal randomized controlled trial. *Computers in Human Behavior* (Mar 2022), 107249. https://doi.org/10.1016/j.chb.2022.107249

Deborah R Billings, Kristin E Schaefer, Jessie Y C Chen, and Peter A Hancock. 2012. Human-robot interaction: developing trust in robots. In *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction - HRI '12.* ACM Press, Boston, Massachusetts, USA, 109–110. https://doi.org/10.1145/2157689.2157709

Stephanie Blackman, Claudine Matlo, Charisse Bobrovitskiy, Ashley Waldoch, Mei Lan Fang, Piper Jackson, Alex Mihailidis, Louise Nygård, Arlene Astell, and Andrew Sixsmith. 2016. Ambient assisted living technologies for aging well: A scoping review. *Journal of Intelligent Systems* 25, 1 (2016), 55–69.

David E Bloom and Dara Lee Luca. 2016. The global demography of aging: Facts, explanations, future. In *Handbook of the economics of population aging.* Elsevier, 3–56.

Ann Bowling and Paul Dieppe. 2005. What is successful ageing and who should define it? *BMJ* 331, 7531 (Dec 2005), 1548. https://doi.org/10.1136/bmj.331.7531.1548

Josiane J J Boyne and Hubertus J M Vrijhoef. 2013. Implementing telemonitoring in heart failure care: barriers from the perspectives of patients, healthcare professionals and healthcare organizations. *Curr. Heart Fail. Rep.* 10, 3 (Sep 2013), 254–261. https://doi.org/10.1007/s11897-013-0140-1 arXiv:23666901

Heidi Bråthen, Harald Maartmann-Moe, and Trenton Wade Schulz. 2019. The Role of Physical Prototyping in Participatory Design with Older Adults. *International Academy, Research and Industry Association (IARIA)* (2019), 141–146. https://www.duo.uio.no/handle/10852/68794

Philipp Brauner and Martina Ziefle. 2021. Social acceptance of serious games for physical and cognitive training in older adults residing in ambient assisted living environments. *Journal of Public Health* (2021), 1–13. https://doi.org/10.1007/s10389-021-01524-y

Elizabeth Broadbent, Ngaire Kerse, Kathryn Peri, Hayley Robinson, Chandimal Jayawardena, Tony Kuo, Chandan Datta, Rebecca Stafford, Haley Butler, Pratyusha Jawalkar, Maddy Amor, Ben Robins, and Bruce MacDonald. 2016. Benefits and problems of health-care robots in aged care settings: A comparison trial. *Australas J Ageing* 35, 1 (Mar 2016), 23–29. https://doi.org/10.1111/ajag.12190

Felix Carros, Isabel Schwaninger, Adrian Preussner, Dave Randall, Rainer Wieching, Geraldine Fitzpatrick, and Volker Wulf. 2022. Care Workers making Use of Robots: Results of a 3 Month Study on Human-Robot-Interaction within a Care Home. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA). ACM, New York, NY, USA, 23 pages. https://doi.org/10.1145/3491102.3517435

Laura L Carstensen. 1992. Social and emotional patterns in adulthood: support for socioemotional selectivity theory. *Psychol. Aging* 7, 3 (Sep 1992), 331–338. https://doi.org/10.1037//0882-7974.7.3.331 arXiv:1388852

Laura L Carstensen, D M Isaacowitz, and S T Charles. 1999. Taking time seriously. A theory of socioemotional selectivity. *Am. Psychol.* 54, 3 (Mar 1999), 165–181. https://doi.org/10.1037//0003-066x.54.3.165 arXiv:10199217

Duncan Case. 1996. Contributions of Journeys away to the definition of home: An empirical study of a dialectical process. *J. Environ. Psychol.* 16, 1 (Mar 1996), 1–15. https://doi.org/10.1006/jevp.1996.0001

Elizabeth Cha, Jodi Forlizzi, and Siddhartha S. Srinivasa. 2015. Robots in the Home: Qualitative and Quantitative Insights into Kitchen Organization. In *HRI '15: Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction.* Association for Computing Machinery, New York, NY, USA, 319–326. https://doi.org/10.1145/2696454.2696465

Mohamed-Amine Choukou, Taylor Shortly, Nicole Leclerc, Derek Freier, Genevieve Lessard, Louise Demers, and Claudine Auger. 2021. Evaluating the acceptance of ambient assisted living technology (AALT) in rehabilitation: A scoping review. *International Journal of Medical Informatics* 150 (2021), 104461. https://doi.org/10.1016/j.ijmedinf.2021.104461

Grazia Cicirelli, Roberto Marani, Antonio Petitti, Annalisa Milella, and Tiziana D'Orazio. 2021. Ambient Assisted Living: A Review of Technologies, Methodologies and Future Perspectives for Healthy Aging of Population. *Sensors* 21, 10 (2021), 3549. https://doi.org/10.3390/s21103549

Carlos A. Cifuentes, Maria J. Pinto, Nathalia Céspedes, and Marcela Múnera. 2020. Social Robots in Therapy and Care. *Curr. Robot. Rep.* 1, 3 (Sep 2020), 59–74. https://doi.org/10.1007/s43154-020-00009-2

Sebastian Cmentowski and Jens Krüger. 2020. Playing With Friends - The Importance of Social Play During the COVID-19 Pandemic. In *Extended Abstracts of the 2020 Annual Symposium on Computer-Human Interaction in Play (Virtual Event, Canada) (CHI PLAY '20).* Association for Computing Machinery, New York, NY, USA, 209–212. https://doi.org/10.1145/3383668.3419911

Luís Correia, Daniel Fuentes, José Ribeiro, Nuno Costa, Arsénio Reis, Carlos Rabadão, João Barroso, and António Pereira. 2021. Usability of Smartbands by the Elderly Population in the Context of Ambient Assisted Living Applications. *Electronics* 10, 14 (2021), 1617. https://doi.org/10.3390/electronics10141617

S Dahlin-Ivanoff, M Haak, A Fänge, and S Iwarsson. 2007. The multiple meaning of home as experienced by very old Swedish people. *Scand. J. Occup. Ther.* 14, 1 (2007), 25–32. https://doi.org/10.1080/11038120601151714 arXiv:17366075

Maartje de Graaf, Somaya Ben Allouch, and Jan van Dijk. 2017. Why Do They Refuse to Use My Robot? Reasons for Non-Use Derived from a Long-Term Home Study. In *HRI '17: Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction.* Association for Computing Machinery, New York, NY, USA, 224–233. https://doi.org/10.1145/2909824.3020236

Maartje M A de Graaf, Bertram F Malle, Anca Dragan, and Tom Ziemke. 2018. Explainable Robotic Systems. In *Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction (Chicago, IL, USA) (HRI '18).* ACM, New York, NY, USA, 387–388. https://doi.org/10.1145/3173386.3173568

Maartje M A de Graaf, Somaya Ben Allouch, and Tineke Klamer. 2015. Sharing a life with Harvey: Exploring the acceptance of and relationship-building with a social robot. *Computers in Human Behavior* 43 (Feb 2015), 1–14. https://doi.org/10.1016/j.chb.2014.10.030

Jessica Diggs. 2008. Activity Theory of Aging. In *Encyclopedia of Aging and Public Health.* Springer, Boston, MA, Boston, MA, USA, 79–81. https://doi.org/10.1007/978-0-387-33754-8

Ha Manh Do, Minh Pham, Weihua Sheng, Dan Yang, and Meiqin Liu. 2018. RiSH: A robot-integrated smart home for elderly care. *Rob. Auton. Syst.* 101 (Mar 2018), 74–92. https://doi.org/10. 1016/j.robot.2017.12.008

Lucile Dupuy, Charles Consel, and Hélène Sauzéon. 2016. Self determination-based design to achieve acceptance of assisted living technologies for older adults. *Computers in Human Behavior* 65 (2016), 508–521.

Erik H Erikson and Joan M Erikson. 1998. *The Life Cycle Completed (Extended Version): A Review.* W. W. Norton & Company.

David Feil-Seifer and Maja J Matarić. 2009. Human RobotInteraction. In *Encyclopedia of Complexity and Systems Science.* Springer, New York, NY, New York, NY, USA, 4643–4659. https://doi. org/10.1007/978-0-387-30440-3

Shira H Fischer, Daniel David, Bradley H Crotty, Meghan Dierks, and Charles Safran. 2014. Acceptance and Use of Health Information Technology By Community-Dwelling Elders. *Int. J. Med. Inf.* 83, 9 (Sep 2014), 624. https://doi.org/10.1016/j.ijmedinf.2014.06.005

Geraldine Fitzpatrick. 2003. *Locales Framework: Understanding and Designing for Wicked Problems.* Kluwer Academic Publishers, Norwell, MA, USA.

Geraldine Fitzpatrick and Gunnar Ellingsen. 2013. A Review of 25 Years of CSCW Research in Healthcare: Contributions, Challenges and Future Agendas. *Computer Supported Cooperative Work* 22, 4 (Aug 2013), 609–665. https://doi.org/10.1007/s10606-012-9168-0

Geraldine Fitzpatrick, Alina Huldtgren, Lone Malmborg, Dave Harley, and Wijnand Ijsselsteijn. 2015. Design for Agency, Adaptivity and Reciprocity: Reimagining AAL and Telecare Agendas. In *Designing Socially Embedded Technologies in the Real-World,* Volker Wulf, Kjeld Schmidt, and David Randall (Eds.). Springer London, London, 305–338. https://doi.org/10.1007/978-1-4471-6720-4_13

Chiara Foà, Maria Cristina Guarnieri, Giorgia Bastoni, Barbara Benini, Olimpia Maria Giunti, Manola Mazzotti, Cristina Rossi, Alessandra Savoia, Leopoldo Sarli, and Giovanna Artioli. 2020. Job satisfaction, work engagement and stress/burnout of elderly care staff: a qualitative research. *Acta Biomed.* 91, Suppl 12 (2020). https://doi.org/10.23750/abm.v91i12-S.10918

Liam Foster and Alan Walker. 2015. Active and successful aging: a European policy perspective. *Gerontologist* 55, 1 (Feb 2015), 83–90. https://doi.org/10.1093/geront/gnu028 arXiv:24846882

Susanne Frennert, Håkan Eftring, and Britt Östlund. 2013. Older People's Involvement in the Development of a Social Assistive Robot. In *Proceedings of the 5th International Conference on Social Robotics - Volume 8239 (Bristol, UK) (ICSR 2013).* Springer-Verlag, Berlin, Heidelberg, 8–18. https://doi.org/10.1007/978-3-319-02675-6_2

Susanne Frennert, Britt Östlund, and Håkan Eftring. 2012. Capturing seniors' requirements for assistive robots by the use of attention cards. *NordiCHI 2012: Making Sense Through Design - Proceedings of the 7th Nordic Conference on Human-Computer Interaction* (Oct 2012), 783–784. https://doi.org/10.1145/2399016.2399146

Vera Gallistl, Alexander Seifert, and Franz Kolland. 2021. COVID-19 as a "Digital Push?" Research Experiences From Long-Term Care and Recommendations for the Post-pandemic Era. *Front Public Health* 9 (May 2021). https://doi.org/10.3389/fpubh.2021.660064

Eva Ganglbauer, Geraldine Fitzpatrick, and Rob Comber. 2013. Negotiating food waste: Using a practice lens to inform design. *ACM Trans. Comput.-Hum. Interact.* 20, 2 (May 2013), 1–25. https://doi.org/10.1145/2463579.2463582

Theodoros Georgiou, Lynne Baillie, Martin K Ross, and Frank Broz. 2020. Applying the Participatory Design Workshop Method to Explore how Socially Assistive Robots Could Assist Stroke Survivors. In *HRI '20: Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction.* Association for Computing Machinery, New York, NY, USA, 203–205. https://doi.org/10.1145/3371382.3378232

K Gerling, D Hebesberger, C Dondrup, T Körtner, and M Hanheide. 2016. Robot deployment in long-term care: Case study on using a mobile robot to support physiotherapy. *Zeitschrift für Gerontologie und Geriatrie* 49, 4 (Jun 2016), 288–297. https://doi.org/10.1007/s00391-016-1065-6 arXiv:27259706

Catharina Gillsjö, Donna Schwartz-Barcott, and Iréne von Post. 2011. Home: The place the older adult can not imagine living without. *BMC Geriatr.* 11, 1 (Dec 2011), 1–10. https://doi.org/10. 1186/1471-2318-11-10

Peter Griffiths, Chiara Dall'Ora, Michael Simon, Jane Ball, Rikard Lindqvist, Anne-Marie Rafferty, Lisette Schoonhoven, Carol Tishelman, and Linda H Aiken. 2014. Nurses' Shift Length and Overtime Working in 12 European Countries: The Association With Perceived Quality of Care and Patient Safety. *Med. Care* 52, 11 (Nov 2014), 975–981. https://doi.org/10.1097/MLR.0000000000000233

Dave Harley. 2011. *Older people's appropriation of computers and the Internet.* Ph. D. Dissertation. http://sro.sussex.ac.uk/id/eprint/7602

Jean D Hallewell Haslwanter and Geraldine Fitzpatrick. 2017. Issues in the Development of AAL Systems: What Experts Think. In *Proceedings of the 10th International Conference on PErvasive Technologies Related to Assistive Environments (Island of Rhodes, Greece) (PETRA '17)*. ACM, New York, NY, USA, 201–208. https://doi.org/10.1145/3056540.3056554

Jean D Hallewell Haslwanter, Katja Neureiter, and Markus Garschall. 2020. User-centered design in AAL. *Univ. Access Inf. Soc.* 19, 1 (Mar 2020), 57–67. https://doi.org/10.1007/s10209-018-0626-4

Eva Hornecker, Andreas Bischof, Philipp Graf, Lena Franzkowiak, and Norbert Krüger. 2020. The Interactive Enactment of Care Technologies and its Implications for Human-Robot-Interaction in Care. In *NordiCHI '20: Proceedings of the 11th Nordic Conference on Human-Computer Interaction: Shaping Experiences, Shaping Society.* Association for Computing Machinery, New York, NY, USA, 1–11. https://doi.org/10.1145/3419249.3420103

Dominik Hornung, Claudia Müller, Alexander Boden, and Martin Stein. 2016. Autonomy Support for Elderly People through Everyday Life Gadgets. In *GROUP '16: Proceedings of the 19th International Conference on Supporting Group Work.* Association for Computing Machinery, New York, NY, USA, 421–424. https://doi.org/10.1145/2957276.2996284

Bahar Irfan, Aditi Ramachandran, Samuel Spaulding, Dylan F Glas, Iolanda Leite, and Kheng Lee Koay. 2019. Personalization in long-term human-robot interaction. In *HRI '19: Proceedings of the 14th ACM/IEEE International Conference on Human-Robot Interaction.* IEEE Press, 685–686. https://doi.org/10.5555/3378680.3378853

Bahar Irfan, Aditi Ramachandran, Samuel Spaulding, Sinan Kalkan, German I Parisi, and Hatice Gunes. 2021. Lifelong Learning and Personalization in Long-Term Human-Robot Interaction (LEAP-HRI). In *HRI '21 Companion: Companion of the 2021 ACM/IEEE*

International Conference on Human-Robot Interaction. Association for Computing Machinery, New York, NY, USA, 724–727. https://doi.org/10.1145/3434074.3444881

Rose-Marie Johansson-Pajala, Kirsten Thommes, Julia A Hoppe, Outi Tuisku, Lea Hennala, Satu Pekkarinen, Helinä Melkas, and Christine Gustafsson. 2020. Care Robot Orientation: What, Who and How? Potential Users' Perceptions. *Int. J. Social Rob.* 12, 5 (Nov 2020), 1103–1117. https://doi.org/10.1007/s12369-020-00619-y

David O Johnson, Raymond H Cuijpers, James F Juola, Elena Torta, Mikhail Simonov, Antonella Frisiello, Marco Bazzani, Wenjie Yan, Cornelius Weber, Stefan Wermter, Nils Meins, Johannes Oberzaucher, Paul Panek, Georg Edelmayer, Peter Mayer, and Christian Beck. 2014. Socially Assistive Robots: A Comprehensive Approach to Extending Independent Living. *Int. J. Social Rob.* 6, 2 (Apr 2014), 195–211. https://doi.org/10.1007/s12369-013-0217-8

Kelly Joyce and Meika Loe. 2010. A sociological approach to ageing, technology and health. *Sociol. Health Illn.* 32, 2 (Feb 2010), 171–180. https://doi.org/10.1111/j.1467-9566.2009.01219.x

Norisan Abd Karim, Haryani Haron, Wan Adilah Wan Adnan, and Natrah Abdullah. 2018. Dimensions for Productive Ageing. In *Recent Trends in Information and Communication Technology.* Springer, Cham, Switzerland, 781–788. https://doi.org/10.1007/978-3-319-59427-9_80

Tineke Klamer and Somaya Ben Allouch. 2010. Acceptance and use of a social robot by elderly users in a domestic environment. In *4th International ICST Conference on Pervasive Computing Technologies for Healthcare.* IEEE. https://doi.org/10.4108/ICST.PERVASIVEHEALTH2010.8892

Bran Knowles and Vicki L Hanson. 2018. Older Adults' Deployment of 'Distrust'. *ACM Trans. Comput.- Hum. Interact.* 25, 4, Article 21 (Aug. 2018), 25 pages. https://doi.org/10.1145/3196490

H I Krebs, B T Volpe, M L Aisen, and N Hogan. 2021. Increasing productivity and quality of care: robot-aided neuro-rehabilitation. *Journal of rehabilitation research and development.* 37, 6 (Oct 2021), 639–652. https://pubmed.ncbi.nlm.nih.gov/11321000 [Online; accessed 15. Oct. 2021].

Kari Kuutti and Liam J Bannon. 2014. The Turn to Practice in HCI: Towards a Research Agenda. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (Toronto, Ontario, Canada) (CHI '14).* ACM, New York, NY, USA, 3543–3552. https://doi.org/10.1145/ 2556288.2557111

Karine Lan Hing Ting, Mustapha Derras, and Dimitri Voilmy. 2018. Designing human-robot interaction for dependent elderlies: a Living Lab approach. *BCS Learning and Development Ltd. Proceedings of British HCI 2018* (Jan 2018). https://doi.org/10.14236/ewic/HCI2018.142

Joe Langley, Gemma Wheeler, Rebecca Partridge, Remi Bec, Dan Wolstenholme, and Lise Sproson. 2019. Designing with and for Older People. In *Design of Assistive Technology for Ageing Populations.* Springer, Cham, Switzerland, 3–19. https://doi.org/10.1007/978-3-030-26292-1_1

Hee Rin Lee, Selma Šabanović, Wan-Ling Chang, Shinichi Nagata, Jennifer Piatt, Casey Bennett, and David Hakken. 2017b. Steps Toward Participatory Design of Social Robots: Mutual Learning with Older Adults with Depression. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction (Vienna, Austria) (HRI '17).* ACM, New York, NY, USA, 244–253. https://doi.org/10.1145/2909824.3020237

Hee Rin Lee, Selma Šabanović, and Sonya S Kwak. 2017a. Collaborative Map Making: A Reflexive Method for Understanding Matters of Concern in Design Research. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (Denver, Colorado, USA) (CHI '17).* ACM, New York, NY, USA, 5678–5689. https://doi.org/10.1145/3025453.3025535

Katherine H Leith. 2006. "Home is where the heart is…or is it?": A phenomenological exploration of the meaning of home for older women in congregate housing. *Journal of Aging Studies* 20, 4 (Dec 2006), 317–333. https://doi.org/10.1016/j.jaging.2005.12.002

Lili Liu, Eleni Stroulia, Ioanis Nikolaidis, Antonio Miguel-Cruz, and Adriana Rios Rincon. 2016. Smart homes and home health monitoring technologies for older adults: A systematic review. *International Journal of Medical Informatics* 91 (2016), 44–59.

Maxime Lussier, Aline Aboujaoudé, Mélanie Couture, Maxim Moreau, Catherine Laliberté, Sylvain Giroux, Hélène Pigot, Sébastien Gaboury, Kévin Bouchard, Patricia Belchior, Carolina Bottari, Guy Paré, Charles Consel, and Nathalie Bier. 2020. Using Ambient Assisted Living to Monitor Older Adults With Alzheimer Disease: Single-Case Study to Validate the Monitoring Report. *JMIR Med. Inform.* 8, 11 (Nov 2020), e20215. https://doi.org/10.2196/20215 arXiv:33185555

Hannah R Marston, Loredana Ivan, Mireia Fernández-Ardévol, Andrea Rosales Climent, Madelín Gómez-León, Daniel Blanche-T, Sarah Earle, Pei-Chun Ko, Sophie Colas, Burcu Bilir, Halime Öztürk Çalikoglu, Hasan Arslan, Rubal Kanozia, Ulla Kriebernegg, Franziska Großschädl, Felix Reer, Thorsten Quandt, Sandra C Buttigieg, Paula A Silva, Vera Gallistl, and Rebekka Rohner. 2020. COVID-19: Technology, Social Connections, Loneliness, and Leisure Activities: An International Study Protocol. *Front. Sociol.* (2020). 5:574811. doi: 10.3389/fsoc.2020.574811

Ester Martinez-Martin and Angel P del Pobil. 2017. Personal Robot Assistants for Elderly Care: An Overview. In *Personal Assistants: Emerging Computational Technologies.* Springer, Cham, Switzerland, 77–91. https://doi.org/10.1007/978-3-319-62530-0_5

Ruth M Masterson Creber, Kathleen T Hickey, and Mathew S Maurer. 2016. Gerontechnologies for older patients with heart failure: What is the role of smartphones, tablets, and remote monitoring devices in improving symptom monitoring and self-care management? *Current Cardiovascular Risk Reports* 10, 10 (2016). https://doi.org/10.1007/s12170-016-0511-8

Angela Mastrianni, Leah Kulp, and Aleksandra Sarcevic. 2021. Transitioning to Remote User Centered Design Activities in the Emergency Medical Field During a Pandemic. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems (Yokohama, Japan) (CHI EA '21).* Association for Computing Machinery, New York, NY, USA, Article 41, 8 pages. https://doi.org/10.1145/3411763.3443444

D Harrison Mcknight, Michelle Carter, Jason Bennett Thatcher, and Paul F Clay. 2011. Trust in a Specific Technology: An Investigation of Its Components and Measures. *ACM Trans. Manage. Inf. Syst.* 2, 2, Article 12 (July 2011), 25 pages. https://doi.org/10.1145/1985347.1985353

Sanika Moharana, Alejandro E Panduro, Hee Rin Lee, and Laurel D Riek. 2019. Robots for joy, robots for sorrow: community based robot design for dementia caregivers. In *HRI '19: Proceedings of the 14th ACM/IEEE International Conference on Human-Robot Interaction.* IEEE Press, 458–467. https://dl.acm.org/doi/10.5555/3378680.3378757

Jeanne Moore. 2000. Placing Home in Context. *J. Environ. Psychol.* 20, 3 (Sep 2000), 207–217. https://doi.org/10.1006/jevp.2000.0178

Maria Y Nilsson, Stefan Andersson, Lennart Magnusson, and Elizabeth Hanson. 2021. Ambient assisted living technology-mediated interventions for older people and their informal carers in the context of healthy ageing: A scoping review. *Health Science Reports* 4, 1 (2021), e225. https://doi.org/10.1002/hsr2.225

Amara Callistus Nwosu, Bethany Sturgeon, Tamsin McGlinchey, Christian D G Goodwin, Ardhendu Behera, Stephen Mason, Sarah Stanley, and Terry R Payne. 2019. Robotic technology for palliative and supportive care: Strengths, weaknesses, opportunities and threats. *Palliat. Med.* 33, 8 (Jun 2019), 1106–1113. https://doi.org/10.1177/0269216319857628

Gabriel A Orenstein and Lindsay Lewis. 2020. Eriksons Stages of Psychosocial Development. *In StatPearls [Internet].* StatPearls Publishing. https://www.ncbi.nlm.nih.gov/books/NBK556096

Erdman Palmore. 1999. *Ageism: Negative and Positive, 2nd Edition.* Springer Publishing Company.

Katinka E Pani-Harreman, Gerrie J J W Bours, Inés Zander, Gertrudis I J M Kempen, and Joop M A van Duren. 2021. Definitions, key themes and aspects of 'ageing in place': a scoping review. *Ageing & Society* 41, 9 (Sept. 2021), 2026–2059. https://doi.org/10.1017/S0144686X20000094

Sebastiaan T Peek, Sil Aarts, and Eveline J Wouters. 2015. Can smart home technology deliver on the promise of independent living? A critical reflection based on the perspectives of older adults. *Handbook of smart homes, health care and well-being* (2015), 203–214.

David Portugal, Paulo Alvito, Eleni Christodoulou, George Samaras, and Jorge Dias. 2019. A Study on the Deployment of a Service Robot in an Elderly Care Center. *Int. J. Social Rob.* 11, 2 (Apr 2019), 317–341. https://doi.org/10.1007/s12369-018-0492-5

Alejandra Recio-Saucedo, Chiara Dall'Ora, Antonello Maruotti, Jane Ball, Jim Briggs, Paul Meredith, Oliver C Redfern, Caroline Kovacs, David Prytherch, Gary B Smith, and Peter Griffiths. 2018. What impact does nursing care left undone have on patient outcomes? Review of the literature. *J. Clin. Nurs.* 27, 11-12 (Jun 2018), 2248–2259. https://doi.org/10.1111/jocn.14058

Sylvie Renaut, Jim Ogg, Ségolène Petite, and Aline Chamahian. 2015. Home environments and adaptations in the context of ageing. *Ageing & Society* 35, 6 (Jul 2015), 1278–1303. https://doi.org/10.1017/S0144686X14000221

Wendy A Rogers, Travis Kadylak, and Megan A Bayles. 2021. Maximizing the Benefits of Participatory Design for Human–Robot Interaction Research With Older Adults. *Human Factors*, 64(3), 441–450. https://doi.org/10.1177/00187208211037465

Astrid Marieke Rosenthal-von der Pütten, Birgit Lugrin, Sophia C Steinhaeusser, and Lina Klass. 2020. Context Matters! Identifying Social Context Factors and Assessing Their Relevance for a Socially Assistive Robot. In *HRI '20: Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction.* Association for Computing Machinery, New York, NY, USA, 409–411. https://doi.org/10.1145/3371382.3378370

John W Rowe and Robert L Kahn. 1997. Successful aging. *Gerontologist* 37, 4 (1997), 433–440. https://doi.org/10.1093/geront/37.4.433 arXiv:9279031

Kristin E Schaefer. 2016. Measuring Trust in Human Robot Interactions: Development of the "Trust Perception Scale-HRI". In *Robust Intelligence and Trust in Autonomous Systems*, Mittu, R., Sofge, D., Wagner, A., Lawless, W. (eds), (2016). Springer, Boston, MA. https://doi.org/10.1007/978-1-4899-7668-0_10

Kjeld Schmidt. 2018. Practice and technology: on the conceptual foundations of practice-centered computing. In *Socio-informatics: A Practice-based Perspective on the Design and Use of IT Artifacts.* Oxford University Press, 47–103.

Eva-Maria Schomakers, Julia Offermann-van Heek, and Martina Ziefle. 2018. Playfully assessing the acceptance and choice of ambient assisted living technologies by older adults. In *International Conference on Information and Communication Technologies for Ageing Well and e-Health.* Springer, 26–44.

Isabel Schwaninger, Felix Carros, Astrid Weiss, Volker Wulf, and Geraldine Fitzpatrick. 2022. Video connecting families and social robots: from ideas to practices putting technology to work. *Univ. Access Inf. Soc.* (July 2022), 1–13. https://doi.org/10.1007/s10209-022-00901-y

Isabel Schwaninger, Geraldine Fitzpatrick, and Astrid Weiss. 2019. Exploring Trust in Human-Agent Collaboration. *Proceedings of the 17th European Conference on Computer-Supported Cooperative Work: The International Venue on Practice-centred Computing an the Design of Cooperation Technologies - Exploratory Papers.* https://doi.org/10.18420/ecscw2019_ep08

Isabel Schwaninger, Christopher Frauenberger, and Geraldine Fitzpatrick. 2020. Unpacking Forms of Relatedness around Older People and Telecare. *Companion Publication of the 2020 ACM Designing Interactive Systems Conference* (Jul 2020), 163–169. https://doi.org/10.1145/3393914.3395867

Isabel Schwaninger, Florian Güldenpfennig, Astrid Weiss, and Geraldine Fitzpatrick. 2021. What Do You Mean by Trust? Establishing Shared Meaning in Interdisciplinary Design for Assistive Technology. *Int. J. Social Rob.* (Jun 2021), 1–19. https://doi.org/10.1007/s12369-020-00742-w

Andrew J Sixsmith. 2000. An evaluation of an intelligent home monitoring system. *J. Telemed. Telecare* 6, 2 (Apr 2000), 63–72. https://doi.org/10.1258/1357633001935059

Judith Sixsmith, Andrew Sixsmith, Agneta Malmgren Fänge, Dörte Naumann, Csaba Kucsera, Signe Tomsone, Maria Haak, Synneve Dahlin-Ivanoff, and Ryan Woolrych. 2014. Healthy ageing and home: The perspectives of very old people in five European countries. *Soc. Sci. Med*. 106 (Apr 2014), 1–9. https://doi.org/10.1016/j.socscimed.2014.01.006

Rachel E Stuck and Wendy A Rogers. 2018. Older Adults' Perceptions of Supporting Factors of Trust in a Robot Care Provider. *Journal of Robotics* 2018 (Apr 2018). https://doi.org/10.1155/2018/6519713

Lars Tornstam. 2005. *Gerotranscendence: A Developmental Theory of Positive Aging.* Springer Publishing Company.

Christiana Tsiourti, Emilie Joly, Cindy Wings, Maher Ben Moussa, and Katarzyna Wac. 2014. Virtual Assistive Companions for Older Adults: Qualitative Field Study and Design Implications. In *Proceedings of the 8th International Conference on Pervasive Computing Technologies for Healthcare (Oldenburg, Germany) (PervasiveHealth '14).* ICST (Institute for Computer Sciences, SocialInformatics and Telecommunications Engineering), ICST, Brussels, Belgium, Belgium, 57–64. https://doi.org/10.4108/icst.pervasivehealth.2014.254943

Riitta Turjamaa, Aki Pehkonen, and Mari Kangasniemi. 2019. How smart homes are used to support older people: An integrative review. *International Journal of Older People Nursing* 14, 4 (Dec. 2019), e12260. https://doi.org/10.1111/opn.12260

Kenneth J Turner and Marilyn R McGee-Lennon. 2013. Advances in telecare over the past 10 years. *SHTT* 1 (Nov 2013), 21–34. https://doi.org/10.2147/SHTT.S42674

Michel Vacher, Sybille Caffiau, François Portet, Brigitte Meillon, Camille Roux, Elena Elias, Benjamin Lecouteux, and Pedro Chahuara. 2015. Evaluation of a context-aware voice interface for Ambient Assisted Living: Qualitative user study vs. quantitative system evaluation. *ACM Transactions on Accessible Computing* 7, 2 (July 2015), 1–36. https://doi.org/10.1145/2738047

Hanna M Van Dijk, Jane M Cramm, Job Van Exel, and Anna P Nieboer. 2015. The ideal neighbourhood for ageing in place as perceived by frail and non-frail community-dwelling older people. *Ageing & Society* 35, 8 (Sept. 2015), 1771–1795. https://doi.org/10.1017/S0144686X14000622

Vivian Vimarlund, Elizabeth M Borycki, Andre W Kushniruk, and Kerstin Avenberg. 2021. Ambient assisted living: Identifying new challenges and needs for digital technologies and service innovation. *Yearbook of Medical Informatics* 2021 (2021), 141–149. https://doi.org/10.1055/s-0041-1726492

Markus Vincze, Astrid Weiss, Lara Lammer, Andreas Huber, and Gerald Gatterer. 2014. On the discrepancy between present service robots and older persons' needs. In *23rd IEEE international symposium on robot and human interactive communication (IEEE RO-MAN 2014)*.

Kazuyoshi Wada, Takanori Shibata, Tomoko Saito, Kayoko Sakamoto, and Kazuo Tanie. 2005. Psychological and Social Effects of One Year Robot Assisted Activity on Elderly People at a Health Service Facility for the Aged. *Proceedings of the 2005 IEEE International Conference on Robotics and Automation* (2005), 2785–2790.

Kazuyoshi Wada, Takanori Shibata, Tomoko Saito, and Kazuo Tanie. 2004. Effects of robot-assisted activity for elderly people and nurses at a day service center. *Proc. IEEE* 92 (2004), 1780–1788. Doi: 10.1109/JPROC.2004.835378.

Astrid Weiss and Katta Spiel. 2021. Robots beyond Science Fiction: mutual learning in human–robot interaction on the way to participatory approaches. *AI & SOCIETY* 5 (Apr 2021). https://doi.org/10.1007/s00146-021-01209-w

Franz Werner, Sabine Payr, and Katharina Werner. 2015. Potential of Robotics for Ambient Assisted Living: Final Report. Abschlussbericht der Studie Potenzial und Grenzen von aktueller Robotik zur Nutzung im Themenfeld des Ambient Assisted Living. https://iktderzukunft.at/resources/pdf/potential-of-robotics-for-ambient-assisted-living-final-report.pdf

World Health Organization. 2002. Active ageing: a policy framework. https://apps.who.int/iris/handle/10665/67215. [Online; accessed 10. Apr. 2022].

World Health Organization. 2018. Ageing and health. https://www.who.int/news-room/fact-sheets/detail/ageing-and-health. [Online; accessed 10. Apr. 2022].

World Health Organization. 2020a. Ageing: Healthy ageing and functional ability. https://www.who.int/westernpacific/news/q-a-detail/ageing-healthy-ageing-and-functional-ability. [Online; accessed 18. Oct. 2021].

World Health Organization. 2020b. WHO announces COVID-19 outbreak a pandemic. https://www.euro.who.int/en/health-topics/health-emergencies/coronavirus-covid-19/news/news/2020/3/who-announces-covid-19-outbreak-a-pandemic. [Online; accessed 19. Mar. 2021].

Volker Wulf. 2009. Theorien sozialer Praktiken zur Fundierung der Wirtschaftsinformatik. In *Wissenschaftstheorie und gestaltungsorientierte Wirtschaftsinformatik,* Becker J, Krcmar H, and Niehaves, B (eds). (2009). Physica-Verlag HD. 211–224. https://doi.org/10.1007/978-3-7908-2336-3_11

Volker Wulf, Markus Rohde, Volkmar Pipek, and Gunnar Stevens. 2011. Engaging with Practices: Design Case Studies As a Research Framework in CSCW. In *Proceedings of the ACM 2011 Conference on Computer Supported Cooperative Work (Hangzhou, China) (CSCW '11).* ACM, New York, NY, USA, 505–512. https://doi.org/10.1145/1958824.1958902

Agency in Sociotechnical Systems: How to Enact Human–Robot Collaboration

Setareh Zafari, Sabine T. Koeszegi

Abstract

As artificial agents are introduced into diverse workplaces, basic configurations underlying the organization of work undergo a fundamental change. This implies that the work we do is subject to alteration along with who does the work that opens new social challenges. Questions regarding the extent of acceptance of these agents in work settings as well as the consequences of collaborating with artificial agents on human agents indicate the need to better understand the mechanisms that underpin a collaborative sociotechnical system. This book chapter discusses how the interplay between humans and artificial agents enables human–robot collaboration as a new way of working. Thus, we first focused on the agents and their interactive processes in the system to analyze how agency is ascribed to nonhuman entities. Thereafter, the results of two experiments are presented to reflect on the impact of attributing agency to an artificial agent on humans. This study provides recommendations for the design of artificial agents and organizational strategies in terms of which social practices and changes in the working context are required to provide possibilities for successful collaborations.

Keywords

Artificial agents, responsibility attribution, human–robot interaction, new ways of working

1 Agency in Sociotechnical Systems: How to Enact Human–robot Collaboration

Over the last decade, advances in the field of artificial intelligence (AI) have enabled passive objects to become active and agentic. Automation, which was principally focused on physical functions, has now begun to impact cognitive functions, such as complex motor coordination, perception processing, and decision-making [Hollnagel 1995]. An increasing use of assistant systems such as social robots and cobots, which are a specific form of robots that can work closely with humans [Faccio et al. 2019], is an example of artificial agents penetrating our everyday lives. Although a lot of the technology that allows these agents to initiate or respond to a variety of interactions with humans already exists, the social implications of these interactions enter new potentialities that have not been fully addressed.

Extensive research has shown that the subjective experience and willingness of humans to accept this integration is as relevant as the objective properties and functionalities of these technologies [De Graaf and Allouch 2013; Echterhoff et al. 2006]. Considering that humans are integral parts of these systems as sense-making actors and end-users, this book chapter analyzes the challenges associated with the integration of artificial agents into human social systems to facilitate collaboration.

https://doi.org/10.34727/2022/isbn.978-3-85448-052-5_10

Developments in the field of AI suggest an increase in the autonomy of machines. One important implication of such autonomy is the ascribed (social) qualities, such as agency, which affects the perceptions and expectations of humans. The unpredictability of the actions undertaken by artificial agents may lead to situations where agency becomes an issue [Weber et al. 2013]. This unpredictability is mainly due to the inadequate understanding of computational mechanisms. Scholars have long debated the impact of attributing agency to machines on the diffusion of responsibility. For instance, Boos et al. [2013] argued that users can only be held accountable if they understand, predict, and influence work processes [Boos et al. 2013]. This approach emphasizes the fit of an actor's accountability demands and their control capabilities while considering the capacities of robots. Although attributing responsibility is the social function of being in control (i.e., a sense of agency) [Frith 2014], current artificial agents cannot be held responsible. Furthermore, understanding the notion of agency is relevant for improving user acceptance in human–machine interactions [Kim 2016; Lee et al. 2015], building trust in these relationships [Engen et al. 2016], and analyzing the ethical implications of smart technologies [Lin et al. 2012]. These arguments warrant a fresh look at artificial agents and their impact on human agents.

To analyze the challenges associated with collaboration with artificial agents, we focused on work settings and investigated how the interplay between humans and artificial agents enables collaboration. As the social is affected by material dimension but also affecting the material dimension [Leonardi 2012; Orlikowski 2009; Zammuto et al. 2007], it is necessary to study how this integration changes the sociotechnical dynamics of organizations. However, establishing the societal consequences of emerging forms of interaction with technology is beyond the scope of this chapter.

This chapter has been organized as follows. First, it defines the notion of agency and provides the framework for further analysis in sociotechnical systems, where humans are supported by technologies. Next, the results of two user experiments are discussed to provide a detailed picture of how collaboration with artificial agents affects humans and their basic needs. Finally, the contributions are summarized, and the directions of future research are discussed.

2 Enacting Human–robot Collaboration

To obtain a balanced emphasis on the social and technical aspects of working conditions, we used the sociotechnical system (STS) approach. This theoretical framework suggests that within organizations, humans (social) and technology (material) continually constitute the features of others. The notion of STS was

developed by Trist and Bamforth in the mid-twentieth century to describe systems comprising a complex interaction between humans, machines, and the environmental aspects of the work system. Previously, the material dimension of organizations was considered as an external discrete input to the study of the social dimensions of organizations. Thus, it undermined the role of the social context in shaping the designs and uses of new technologies over time. However, this framework follows a relational ontology perspective [Law 2004; Barad 2007] and stresses the reciprocal interrelationship and entanglement of humans and technologies that shape technical and social working conditions. Although the social subsystem comprises individuals as members of the organization, the relationship among them, and their social attributes, the technical subsystem comprises the devices, techniques, and skills used by individuals to perform organizational tasks [Leonardi, 2012]. Thus, with a focus on the constitutive effect of the material and social dimensions, the properties of technologies and humans should be considered to explain how new affordances for working are created [Orlikowski 2009; Zammuto et al. 2007].

An underlying premise of this approach is that capacities for action are enacted in practice [Orlikowski and Scott 2008]. As machines become more sophisticated, understanding the agency attributed to these entities and its impact on humans and collaboration becomes even more critical. Agency is regarded as the capacity to act [Gray et al. 2007]. Two abstracted properties of agency are intentionality and autonomy [Bandura 1999; Banks 2019]. Intentionality is characterized by the capacity of an agent to process the contents of the mental state and justify actions or decisions. According to the theory of action [Davidson 1963], an action is intentional when it is caused by certain mental states. Thus, if no patterns of interaction and coordination based on expectations are identified, it is a coincidence and unintended. Autonomy is a combination of two Greek terms, auto (self) and nomos (governance) and is expressed in two dimensions: self-directedness (i.e., free will) and self-sufficiency (i.e., free act) [Bradshaw et al. 2013]. The former describes the agent's capability to take care of itself and create its agenda, while the latter describes the extent to which an agent is independent of external control. Thus, if no contingency or deviation from the set course is involved, an action is determined and preprogrammed.

Focusing on the agency of representative entities in a sociotechnical system can facilitate the development of more robust theories of the interrelationship between humans and artificial agents within a workplace. Moreover, it can potentially inform future strategic objectives for organizations that aim to integrate artificial agents. Therefore, this study analyzed how social and material entities and their agencies are continually coconstructed to enable a new way of working, namely human–robot collaboration. Human–robot collaboration refers to a collaborative

partnership between humans and robots in completing tasks and focuses on coordinating joint activities between them [Ajoudani et al. 2018].

Collaboration can be differentiated from cooperation, where tasks for achieving a common goal are divided among participants, and each agent is responsible for only a part of the problem-solving. Collaboration is characterized by "the mutual engagement of participants in a coordinated effort to solve the problem together" [Roschelle and Teasley 1995, p. 70]. Therefore, collaborations require all agents to jointly engage in the entire task. That is, collaboration employs a complementarity approach and exceeds existing research that mostly substitutes humans with machines.

Studies have shown that human–AI collaboration can outperform a group of humans and sophisticated AI-based systems [Wang et al. 2016; Siegel 2016]. The resulting team success can be attributed to the unique advantages that emerge from combining human and AI capabilities in a compatible manner [Krüger et al. 2017]. Although the strengths of AI lie in analytical decision-making that involves the gathering and processing of large amounts of data, humans are well-versed in flexibility, creativity, and intuitive decision-making, particularly when heuristics are necessary for decision-making in uncertainty [Dragicevic et al. 2020; Jarrahi 2018]. Thus, artificial agents can extend human capabilities in task performance and decision-making.

To build an effective system, one needs to examine how integrating artificial agents reconfigures the main domains of an organization, including the i) division of labor and ii) integration of efforts. The former focuses on how to distribute tasks and decision rights among agents (human or artificial), and the latter elaborates on how to ensure the alignment of the efforts of different agents with the organizational goals. Therefore, studying the agents within this system is the first step to developing a better sense of the sociotechnical development process. However, using a sociotechnical approach to analyze collaboration with artificial agents does not mean categorizing social and technical actors and their actions but rather, showing the conditions of possibilities for these assumed categories or actors to behave in certain ways. Accordingly, it focuses on the flow of social formulations that enact those actions and performances [Hultin 2019]. Thus, we explored how agents (human and artificial) and their properties and identities are continuously performed to enable collaborative work between humans and robots as a new way of working.

Each subsection refers to an original work of the authors conducted as part of the dissertation. First, we focused on the agents and their interactive processes in the system to analyze how agency is ascribed to nonhuman entities (subsection 2.1). Thereafter, the results of two experiments are presented to reflect on

the impact of attributing agency to an artificial agent on humans (subsections 2.2 and 2.3)

2.1. Robots as Artificial Agents

Several theoretical models such as the Actor–Network Theory [Latour 1996] and Double Dance of Agency [Rose and Jones 2005] suggest that ascribing agency is not limited to humans but also nonhuman entities, such as technologies. We differentiated between the agency of humans and that of machines and studied how these types of agencies are interrelated.

Recent developments in the field of AI suggest an increase in the agency of machines, as we assign them roles that were previously filled by humans. However, the unpredictability of the actions undertaken by artificial agents leads to situations where agency becomes an issue [Weber et al. 2013]. For instance, who would be responsible for the harm that is caused by a self-driving car? Considering that humans and machines do not possess the same capabilities [Engen et al. 2016; Rose and Jones 2005], we investigated the concept of agency and sought to comprehend the properties that humans seek when ascribing agency to nonhuman entities, such as robots.

Previous studies discovered different features related to our perception of machine agency, such as adaptability [Franklin and Graesser 1997], purposeful-looking movement [Scholl and Tremoulet 2000], complementary personalities [Lee et al. 2006], and humanlike appearance [Itoh and Inagaki 2004; Lee et al. 2015]. A seminal study in this area is the work of Rose and Turex [2000], which relates the perceived agency of machines to the human tendency toward anthropomorphism and describes machine agency as the extent to which machines are perceived by humans as having autonomy [Rose and Turex 2000].

We incorporated variable dimensions to develop a typology of artificial agents from a theoretical perspective. Typology is a conceptual classification that is mostly used in social, rather than natural sciences [Baily 1994]. It is one of the common styles of theorizing that systematically categorizes specific dimensions and features to create distinct types and profiles [Cornelissen 2017]. Classifying the artificial agents enables a deep and extended analysis of theories in previous studies about (social) agency to reflect on the possible consequences of human interactions with artificial agents on human–human interaction.

Depending on how machines control the input–output cycle and pursue the goal, we conceptualized artificial agents in four types, i) Non-AI marginally autonomous agents, ii) AI marginally autonomous agents, iii) AI semiautonomous agents, and iv) AI pseudoautonomous agents [Zafari and Koeszegi 2018]. A key

distinction among these artificial agents is the extent to which they independently perform tasks. The autonomous consideration of AI marginally autonomous agents lies in their ability to move without human intervention, while AI semiautonomous agents adapt their goal settings because of their self-learning capacities. Responsibility implies autonomy; therefore, artificial agents are exempt from the usual responsibility practices and attribution. Thus, by finding such an artificial agent as the source of a failure or negative outcome, we need to understand how it determined the cause of failure to handle the issue and prevent a repetition. Thus, the responsibility for harm caused by artificial agents will always remain with human agents who initiate or manage the collaboration as the artificial agents are under the authority of the human agent in every step of the process.

The insights gained from this work [Zafari and Koeszegi 2018] may support the notion of collaborative agency [Kuziemsky and Cornett 2013]. Thus, agency does not belong to any actors and can be viewed as social affordance that emerges from the interaction between humans and artificial agents. This correlates with the relational ontology that argued that agency is constantly forming within the action [Law 2004; Barad 2007]. The agency attributed to an agent (human or artificial) may change in scale, over time and from one situation to another. Therefore, emphasis needs to be placed on the large-scale "system" at the heart of the analysis rather than discussing single agents to better elucidate organizational challenges.

2.2. Attitudes toward Artificial Agents

During collaboration, the activities of humans and robots occur in the same physical and social spaces [Dautenhahn and Sanders 2011]. This highlights the importance of the social aspects of interaction between these agents. Furthermore, ascribing agency to another entity highly depends on the physical and behavioral features of the entity and the characteristics of the perceiver [Takayama 2011; Waytz et al. 2010]. Studies on human–robot interaction (HRI) have mostly focused on the former [e.g., Itoh and Inagaki 2004; Lee et al. 2015; Lee et al. 2006], and there is still an extremely limited understanding of the cognitive processes that occur during HRI. Several technological features in robotics (such as increased sensitivity and safety) allow collaborative robots to support joint action in close contact with humans within a shared workspace [Bauer et al. 2008]. Inadequate effective management of social and cognitive features such as psychological safety [Edmondson 1999] and situational awareness [Cramton 2001] burden the collaboration between humans and robots. To provide insight into the cognition and intentional stance of humans while interacting with artificial agents,

it is necessary to analyze the conditions that ensure the acceptance of the support of artificial agents without limiting human agency.

User studies about the Roomba robot showed that the owners exhibited different behaviors from the same robot vacuum cleaner; some gave it a name while it emptied its way, and some treated it like any other home application and did not talk to it [Takayama 2011; Forlizzi and Disalvo 2006]. This implied that the status of an entity's agency is not static, and the predefined and programmed functions of the entity and the perception of agency influence how we behave and interact with an entity. Moreover, recent studies [Appel et al. 2020; Złotowski et al. 2017] have shown that experience, as a dimension of mind perception, as well as agency, is related to an uncanny feeling toward humanlike robots and requires a better understanding of how ascribing agency elicits uncanniness or negative responses.

Although there is minimal theoretical knowledge regarding the agency of robots, it is necessary to not only describe but conduct an empirical study to explain under which conditions attributed agency positively/negatively impacts the attitudes toward robots. Thus, we conducted a vignette study and investigated the mechanism of the attitudes toward artificial agents. Vignettes refer to text, images, or videos that shortly describe a specific situation to evoke the attitudes or beliefs of participants concerning the present situation [Hughes and Huby 2002]. The flexibility of vignettes allows the exploration of factors and elements of interest by combining traditional survey and experimental design [Steiner et al. 2016]. Participants were asked to watch a video and respond to a postvideo questionnaire from the perspective of the vignette character as if they were that person in that situation.

We created two videos in which a human and robot collaborate to assemble a product (Figure 1). The main difference between the conditions is that under the "low agency" condition, the robot's behavior was relatively deterministic, while its behavior under the "high agency" condition was unpredictable. Thus, the actions of the cobot were not presuggested but were imperatively used to reflect the high level of autonomy.

The results showed that attributing high levels of agency to robots was associated with negative attitudes toward them only when individuals perceived low control during collaboration [Zafari and Koeszegi 2020]. Therefore, the lower the levels of decision control (inhibiting human autonomy), the lower the positive attitudes toward the robot with a high level of agency. Although preliminary, this finding highlights the role of the perception of control in promoting positive attitudes toward artificial agents. It implied that people do not perceive the high level of agency for artificial agents as negative except when they feel a lack of control

during the work process. Furthermore, because perceived control is highly related to the diffusion of responsibility [Bandura 1991], it is necessary to consider the nature of perceived control in the collaborative context and establish approaches to enhance the perceptions of control for individuals working alongside artificial agents.

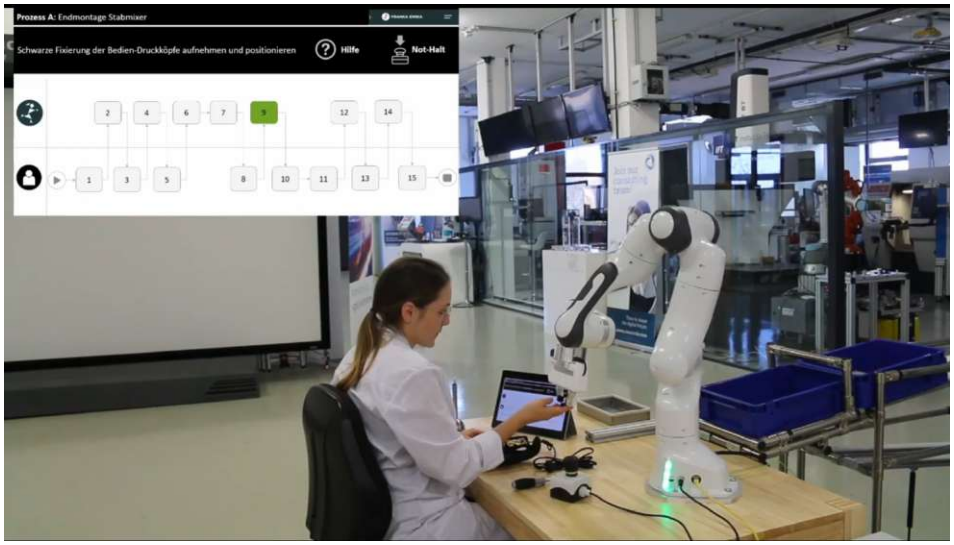


Figure 1 Screenshot of the video vignette that represents an artificial agent collaborating with a human agent

2.3. Interaction Style of Artificial Agents

A previous study on computers-are-social-actors established that social responses to computers fall under natural reactions to social situations; therefore, the principles drawn from sociology and social psychology are relevant for user interface design [Nass et al. 1994]. We interact with others according to our interpretation of the stimulus we receive from them [Blumer 1969]. The interpretation is a flexible social construct, which depends on the context and party involved [Pinch and Bijker 1984]. It helps us to clarify what to expect from the other party and is the basis for our future interactions.

Research predicts that service robots will soon be used within the social sphere of human agents as “natural” interaction partners [Floridi 2008]. With an increase in the entanglement of HRI, questions regarding the needs concerning the design of service robot applications have arisen. The appropriate design and implementation of robots serving with humans have been confirmed to be more challenging than old-fashioned industrial robots serving for humans. Robots serving for humans need to be capable of operating more or less autonomously and learning

from errors, while robots serving with humans require the ability to communicate and interact with humans on a level involving understanding and responsiveness toward the human interaction partner [Kolbeinsson et al. 2019; Decker 2013]. This places a high demand on the quality of the interaction between humans and robots.

Considering that how a robot interacts with people can affect the efficiency of collaboration [Schulz et al. 2018], we focused on the interaction style of artificial agents and conducted a laboratory-based experiment with a Pepper robot developed by SoftBank Robotics using a built-in software. To design the interaction style of the robot, we referred to the "Big Two" dimensions of agency and communion [Bakan 1966]. Although the external validity of laboratory experiments is relatively lower than that of field experiments, they are a common method for HRI studies. A possible explanation for this is that most service robots are not easily accessible for daily usage since they are still in the research and development phase [Von der Puetten et al. 2018]. Laboratory experiments benefit from the high control over the extraneous variables that facilitate the replication of the conditions [Tanner 2018]. Therefore, they are useful for testing predictions and providing implication for designers of future robots.

We created two conditions of "person-oriented" and "task-oriented" interaction styles in which a service robot verbally assisted participants while they were building a house of cards. The robot under the person-oriented condition focused on socioemotional support and provided the participants with simple motivational phrases, while the robot's focus under the task-oriented condition was on task performance and provided guidance concerning the goal and participant's progress.

The experimental results showed that people interacting with a robot with a person-oriented interaction style reported higher self-efficacy in HRI, compared with that of a robot with a task-oriented interaction style. Moreover, we observed that several dimensions of the personality of a robot (specifically, extraversion, agreeableness, and emotional stability) can be simulated via robot verbal and speech interface design [Zafari et al. 2019]. These findings suggest the role of the interaction style of the robot in promoting perceived self-efficacy, which is crucial in developing trust in HRI. This implies that when a robot places emphasis on forming and maintaining a social relationship rather than pursuing goals and manifesting skills, an individual's belief concerning their capabilities to perform in a particular situation heightens.

3 Discussion and Conclusion

Investigating the role of technology in organizations is a continuing concern within organizational research. Although new technologies are embraced for their capacity to create new ways of working, their disruptive impacts should not be undermined. This calls for a social science and human factor perspective to analyze the domains where these technologies potentially can and should be used and where they can but should not be used as their implementation may pose threats and challenges to organizations and society.

This book chapter discusses human–robot collaboration as a representative form of sociotechnical systems. It contributes to a better understanding of the impact of artificial agents on the behavior of human agents by discussing how the successful integration of the emerging technologies of AI and robotics in organizations depends not only on overcoming technical limitations but also considering social challenges. We demonstrated how the integration of artificial agents into social systems is reshaping the organization of work as the engagement with artificial agents creates the conditionality that makes certain practices enacted. Therefore, changes in work organization depend on assumed human agency, and the engagement with artificial agents creates a new arrangement of shared control in which agency is assigned and attributed to humans and artificial agents. This collaboration mindset helps to position human agents as the cocreators of the outcomes rather than the passive receiver of services provided by artificial agents. Thus, to better elucidate organizational challenges, we need to emphasize the system rather than the analysis of single agents.

The empirical findings reported in this chapter shed new light on social processes and their contribution to how people collaborate with artificial agents. As the ascribed agency to robots increases, the use of social processes in HRI also increases [Breazeal 2004]. The artificial nature of these agents presents several implications for their social interactions with humans; therefore, we suggested a set of contextual factors that influence the enactment of human–robot collaboration. We observed that the high perception of autonomy for an artificial agent leads to a lower acceptance and positive attitudes toward them when the level of perceived control for a human agent is low (inhibiting human autonomy). Furthermore, we observed that a robot's interaction style in providing feedback could be considered as a factor affecting self-efficacy in human collaborations. From a self-determination theory perspective, experiencing a sense of efficacy must be accompanied by a sense of autonomy [Deci and Ryan 2000] for intrinsic motivation to flourish, as the former resembles the need for competence, while the latter resembles the need for autonomy. Thus, the approach to the design of artificial agents requires the satisfaction of these human needs.

This emphasizes the importance of informal structure in enhancing the success of technological integration. The impact of delegating decisions and assigning roles to artificial agents in organizations is not limited to formal domains of organization (i.e., division of labor and integration processes) because the basic needs of individuals (i.e., needs for autonomy and competence), their work roles, and the social organizational structure are also affected. These findings suggest that for a successful integration of artificial agents into workspaces, a mindful consideration of the social components of interaction among humans and artificial agents is essential.

In addition to its exploratory nature, this chapter offers insight into which practices and changes in work organization are required to provide possibilities for successful integration. In this process, the key constructs are defined, the relationship between them is elucidated, and findings are discussed to demonstrate the viability of theoretical methods that offer minimal empirical support. This contribution can be classified as an intermediate theory [Edmondson and McManus 2007] that identifies new relationships among phenomena by reconceptualizing explanatory frameworks.

The scope of this study was limited in terms of work organization and analyzed how advances in the field of AI and robotics are affecting collaboration. A natural progression of this study is to analyze the possibilities and consequences of integrating these technologies into the tasks and processes that cannot yet be assigned to artificial agents, such as those requiring creativity. Further research can go beyond dyadic interactions between a human and artificial agent and explore how the team characteristics (such as the diversity or composition of a team) affect work dynamics and collaboration.

The collaboration process has a fundamental social component that robots working as the physical interaction partners of the human agent present a great risk on fundamental structures that are usually brought forth within human–human interaction, e.g., social norms. People expect artificial agents to apply the same norms that govern human–human interaction, and behavior that is not performed sufficiently similar to that of humans hinders the pragmatics of interaction [Sciutti et al. 2015]. Although humans will adapt to the capabilities of artificial agents [Hirschmanner et al. 2021] as well as the functionality of the sociotechnical system [Zafari et al. 2021], the impacts of constant interaction with artificial agents on the development and changes in social norms remain unclear. As Goffman [1983] emphasizes, the social self and individual actor are created through interactions [Goffman 1983]. The societal consequences of artificial agents penetrating the social lives of humans are intriguing and can be explored for further research.

Bibliography

Arash Ajoudani, Andrea Zanchettin, Serena Ivaldi, Alin Albu-Schaffer, Kazuhiro Kosuge, and Oussama Khatib. 2018. Progress and prospects of the human–robot collaboration. *Autonomous Robots* 42, 5 (2018), 957–975.

Markus Appel, Silvana Weber, Stefan Krause, and Martina Mara. 2016. On the eeriness of service robots with emotional capabilities. In *The Eleventh ACM/IEEE International Conference on Human Robot Interaction.* 411–412.

Kenneth Bailey. 1994. *Typologies and taxonomies: An introduction to classification techniques.* 102. Sage.

David Bakan. 1966. *The duality of human existence: Isolation and communion in western man.* Beacon Press.

Albert Bandura. 1991. Social cognitive theory of self-regulation. *Organizational Behavior and Human Decision Processes* 50, 2 (1991), 248–287.

Albert Bandura. 1999. Social cognitive theory: An agentic perspective. *Asian journal of social psychology* 2, 1 (1999), 21-41.

John Banks. 2019. A perceived moral agency scale: Development and validation of a metric for humans and social machines. *Computers in Human Behavior* 90 (2019), 363–371.

Karen Barad. 2007. *Meeting the Universe Halfway: Quantum physics and the entanglement of matter and meaning.* Duke University Press.

Andrea Bauer, Drik Wollherr, and Martin Buss. 2008. Human–robot collaboration: A survey. *International Journal of Humanoid Robotics* 5, 1 (2008), 47-66.

Herbert Blumer. 1969. *Symbolic interactionism: Perspective and method.* Univ. of California Press.

Daniel Boos, Hannes Guenter, Gudela Grote, and Katharina Kinder. 2013. Controllable accountabilities: The Internet of Things and its challenges for organisations. *Behaviour and Information Technology* 32, 5 (2013), 449–467.

Jeffery Bradshaw, Robert Hoffman, David Woods, and Matthew Johnson. 2013. The seven deadly myths of "autonomous systems." *IEEE Intelligent Systems* 28, 3 (2013), 54–61.

Robert Siegel. 2016. 20 Years later, humans still no match for computers on the chessboard. *NPR*. Available at https://www.npr.org/sections/alltechconsidered/2016/10/24/499162905/20-years-later-humans-still-no-match-for-computers-on-the-chessboard

Joep Cornelissen. 2017. Editor's comments: Developing propositions, a process model, or a typology? Addressing the challenges of writing theory without a boilerplate. *Academy of Management Review* 42, 1 (2017), 1-9.

Catherine Cramton. 2001. The mutual knowledge problem and its consequences for dispersed collaboration. *Organization Science* 12, 3 (2001), 346–371.

Kerstin Dautenhahn, and Joe Sanders. 2011. Introduction. In *New frontiers in human-robot interaction,* Dautenhahn, K, and Sanders, J (Eds.). John Benjamins Publishing. 1–5.

Donald Davidson. 1963. Actions, Reasons, and Causes, *Journal of Philosophy* 60, 23 (1963), 685-700.

Edward Deci, and Richard Ryan. 2000. The “What” and “Why” of Goal Pursuits: Human Needs and the Self-Determination of Behavior. *Psychological Inquiry* 11, 4 (2000), 227-268. DOI:10.1207/S15327965PLI1104_01

Maartje De Graaf, and Somaya Ben Allouch. 2013. Exploring influencing variables for the acceptance of social robots. *Robotics and Autonomous Systems* 61, 12 (2013), 1476–1486.

Michael Decker. 2013. Robotik. In *Handbuch Technikethik,* Grunwald, A, Simonidis-Puschmann, M (Eds.). Stuttgart JB Metzler, 354–358, https://doi.org/10.1007/978-3-476-05333-6_67

Gerald Echterhoff, Gerd Bohner, and Frank Siebler. 2006. “Social Robotics” und Mensch-Maschine-Interaktion. *Zeitschrift Für Sozialpsychologie* 37, 4 (2006), 219–231.

Amy Edmondson, and Stacy McManus. 2007. Methodological fit in management field research. *Academy of Management Review* 32, 4 (2007), 1246–1264.

Vegard Engen, J Brian Pickering, and Paul Walland. 2016. Machine agency in human-machine networks; impacts and trust implications. *Lecture Notes in Computer Science*, 9733. Springer, Cham. 96–106.

Maurizio Faccio, Matteo Bottin, and Giulio Rosati. 2019. Collaborative and traditional robotic assembly: A comparison model. *The International Journal of Advanced Manufacturing Technology* 102, 5 (2019), 1355–1372.

Luciano Floridi. 2008. Artificial intelligence’s new frontier: Artificial companions and the fourth revolution. *Metaphilosophy* 39 (2008), 651–655.

Jodi Forlizzi, and Carl DiSalvo. 2006. Service robots in a domestic environment: A study of the Roomba vacuum in the home. In *HRI ‘06: Proceedings of the 1st ACM SIGCHI/SIGART conference on Human-robot interaction*. ACM. https://doi.org/10.1145/1121241.1121286

Stan Franklin, and Art Graesser. 1997. Is It an Agent, or Just a Program?: A Taxonomy for Autonomous Agents. In: Müller J P, Wooldridge M J, Jennings N R (eds) *Intelligent Agents III Agent Theories, Architectures, and Languages. ATAL 1996. Lecture Notes in Computer Science,* vol 1193. Springer, Berlin, Heidelberg. https://doi.org/10.1007/BFb0013570

Chris Frith. 2014. Action, agency and responsibility. *Neuropsychologia* 55, 1 (2014), 137–142.

Heather Gray, Kurt Gray, and Daniel Wegner. 2007. Dimensions of mind perception. *Science* 315, 5812 (2007), 619. https://doi.org/10.1126/science.1134475

Matthias Hirschmanner, Stephanie Gross, Setareh Zafari, Brigitte Krenn, Friedrich Neubarth, and Markus Vincze. 2021. Investigating Transparency Methods in a Robot Word-Learning System and Their Effects on Human Teaching Behaviors. In *2021 30th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN).* 175–182. https://doi.org/10.1126/science.1134475

Erik Hollnagel. 1995. Cognitive functions and automation: principles of human-centred automation. *Advances in Human Factors/Ergonomics* 20 (1995), 971–976.

Rhidian Hughes, and Meg Huby. 2002. The application of vignettes in social and nursing research. *Journal of Advanced Nursing* 37,4 (2002), 382–386.

Lotta Hultin. 2019. Information and organization on becoming a sociomaterial researcher: Exploring epistemological practices grounded in a relational, performative ontology. *Information and Organization* 29, 2 (2019), 91–104.

Makoto Itoh, and Toshiyuki Inagaki. 2004. A microworld approach to identifying issues of human-automation systems design for supporting operator's situation awareness. *International Journal of Human-Computer Interaction* 17, 1 (2004), 3–24.

Mohammad Jarrahi. 2018. Artificial intelligence and the future of work: *Human-AI symbiosis in organizational decision making. Business Horizons* 61, 4 (2018), 577–586.

Katherine Kim. 2016. Interacting socially with the internet of things (IoT): Effects of source attribution and specialization in human-IoT interaction. *Journal of Computer-Mediated Communication* 21,2 (2016), 420–435.

Ari Kolbeinsson, Erik Lagerstedt, and Jessica Lindblom. 2019. Foundation for a classification of collaboration levels for human-robot cooperation in manufacturing human-robot cooperation in manufacturing. *Production and Manufacturing Research* 7, 1 (2019), 448–471.

Craig Kuziemsky, and Janet Cornett. 2013. A model of collaborative agency and common ground. *In ITCH.* IOS Press. 388–392.

Bruno Latour. 1996. On actor-network theory: A few clarifications. *Soziale Welt*, 369-381.

James Law. 2004. *After method: Mess in social science research.* Routledge.

Jae-Gil Lee, Ki Joon Kim, Sangwon Lee, and Dong-Hee Shin. 2015. Can autonomous vehicles be safe and trustworthy? Effects of appearance and autonomy of unmanned driving systems. *International Journal of Human-Computer Interaction,* 31(10), 682–691. https://doi.org/10.1080/10447318.2015.1070547

Kwan Lee, Wei Peng, Seung Jin, and Chang Yan. 2006. Can robots manifest personality?: An empirical test of personality recognition, social responses, and social presence in human-robot interaction. *Journal of Communication* 56, 4 (2006), 754–772.

Paul Leonardi. 2012. Materiality, sociomateriality and socio-technical systems: What do these terms mean? How are they related? Do we need them? In Leonardi, P M, Nardi, B A, Kallinikos, J (Eds.), *Materiality and organizing: Social interaction in a technological world.* Oxford University Press. 25-48.

Patrick Lin, P., Keith Abney, and George Bekey. 2012. *Robot Ethics: The social and ethical implications of robotics.* MIT Press, MA2012.

Clifford Nass, Jonathan Steuer, and Ellen Tauber. 1994. Computers are social actors. *CHI '94: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 72–78. https://doi.org/10.1145/191666.191703

Wanda Orlikowski, and Susan Scott. 2008. Sociomateriality: Challenging the separation of technology, work and organisation. *The Academy of Management Annals* 2, 1 (2008), 433-474.

Wanda Orlikowski. 2009. The sociomateriality of organisational life: considering technology in management research. *Cambridge Journal of Economics* 34, 1 (2009), 125–141.

Trevor Pinch, and Wiebe Bijker. 1984. The social construction of facts and artefacts: Or how the sociology of science and the sociology of technology might benefit each other. *Social Studies of Science* 14, 3 (1984), 399–441.

Jeremy Roschelle, and Stephanie Teasley. 1995. The construction of shared knowledge in collaborative problem solving. In *Computer supported collaborative learning,* Springer. 69–97.

Jeremy Rose, and Matthew Jones. 2005. The double dance of agency: A socio-theoretic account of how machines and humans interact. *An International Journal on Communication, Information Technology and Work* 1, 1 (2005), 19–37.

Brian Scholl, and Patrice Tremoulet. 2000. Perceptual causality and animacy. *Trends in Cognitive Sciences* 4, 8 (2000), 299–309.

Ruth Schulz, Philipp Kratzer, and Marc Toussaint. 2018. Preferred interaction styles for human-robot collaboration vary over tasks with different action types. *Frontiers in Neurorobotics,* 12 (2018), 36.

Alessandra Sciutti, Caterina Ansuini, Cristina Becchio, and Giulio Sandini. 2015. Investigating the ability to read others intentions using humanoid robots. *Frontiers in Psychology,* 6 (2015), 1–6.

Peter Steiner, Christiane Atzmüller, and Dan Su. 2016. Designing valid and reliable vignette experiments for survey research: A case study on the fair gender income gap. *Journal of Methods and Measurement in the Social Sciences,* 7, 2 (2016), 52–94.

Leila Takayama. 2012. Perspectives on agency interacting with and through personal robots. In *Human-computer interaction: the agency perspective.* Springer, Berlin, 195–214

Kerry Tanner, 2018. Experimental research. In *Research Methods.* Chandos Publishing. 337–356

Astrid M von der Pütten, Nicole C Krämer, Cristina Becker-Asano, Kohei Ogawa, Suichi Nishio, and Hiroshi Ishiguro. 2018. At the café—Exploration and analysis of people's nonverbal behavior toward an android. In *Geminoid Studies,* Ishiguro H, Dalla Libera F (eds). Springer, 375–397. https://doi.org/10.1007/978-981-10-8702-8_24

Dayong Wang, Aditya Khosla, Rishab Gargeya, Humayun Irshad, and Andrew Beck. 2016. Deep learning for identifying metastatic breast cancer. *arXiv preprint* arXiv:1606.05718

Adam Waytz, Kurt Gray, Nicholas Epley, and Daniel Wegner. 2010. Causes and consequences of mind perception. *Trends in Cognitive Sciences*, 14, 8 (2010), 383–388.

Rolf Weber, Angela Pereira, Francien Dechesne, Job Timmermans, Rob van Kranenburg, and Hendrik Lehn. 2013. *Fact sheet - Ethics Subgroup IoT - Version 4. 0.* Delft University of Technology Chair Ethics Subgroup IoT Expert Group, 1–21.

Setareh Zafari, Sabine Koeszegi, Michael Filzmoser. 2021. Human Adaption in the Collaboration with Artificial Agents. In *Konnektivität Über die Bedeutung von Zusammenarbeit in der virtuellen Welt*, J Fritz, N Tomaschek (Eds.). Waxmann Verlag GmbH, Münster, Deutschland, (2021), 97–106.

Setareh Zafari and Sabine Koeszegi. 2020. Attitudes toward attributed agency: Role of perceived control. *International Journal of Social Robotics.* 13 (2020), 2071–2080. https://doi.org/10.1007/s12369-020-00672-7

Setareh Zafari, Isabel Schwaninger, Matthias Hirschmanner, Christina Schmidbauer, Astrid Weiss, and Sabine Koeszegi. 2019. "You Are Doing so Great!" – The Effect of a Robot's Interaction Style on Self-Efficacy in HRI. In *proceedings of IEEE International Symposium on Robot and Human Interactive Communication (ROMAN).* pp. 1–7, doi: 10.1109/RO-MAN46459.2019.8956437.

Setareh Zafari and Sabine Koeszegi. 2018. Machine agency in socio-technical systems: A typology of autonomous artificial agents. In *Proceeding of 2018 IEEE Workshop on Advanced Robotics and its Social Impacts (ARSO).* Doi: 10.1109/ARSO.2018.8625765.

Raymond Zammuto, Terri Griffith, Ann Majchrzak, Daniel Dougherty, and Samer Faraj. 2007. Information technology and the changing fabric of organization. *Organization Science* 18, 5 (2007), 749–762

Jacob Złotowski, Kumar Yogeeswaran, and Christoph Bartneck. 2017. Can we control it? Autonomous robots threaten human identity, uniqueness, safety, and resources. *International Journal of Human-Computer Studies.* 100(2017), 48–54.

Adaptive Task Sharing between Humans and Collaborative Robots in a Manufacturing Environment

Christina Schmidbauer, Sebastian Schlund

Abstract

Collaborative robots (referred to as cobots) can have a significant potential impact on manufacturing processes by enabling new task allocation possibilities, resulting in improved economic efficiency and human factors/ergonomics. In this chapter, a method for sharing tasks adaptively between humans and cobots is designed, developed, demonstrated, and evaluated. State-of-the-art task allocation approaches and their shortcomings regarding flexibility and human factors/ergonomics are presented. The three parts of the proposed adaptive task sharing method, i.e., task analysis, assignment, and visualization, are specifically described. Also, case studies are demonstrated, and the obtained results are evaluated. A discussion, conclusion, and the research outlook on human–robot interaction in future manufacturing processes conclude this chapter.

Keywords

Collaborative Robots, Flexible Production Systems, Manufacturing, Task Allocation, Task Sharing

1 Introduction

Collaborative robotic arms, referred to as cobots, entered the market more than 20 years ago. Their design differs from that of conventional industrial robots because the objective is to ensure a safe interaction between cobots and human workers [Albu-Schäffer et al. 2007]. Specifically, these robots are equipped with inherent safety measures and intuitive user interfaces. Therefore, after a short period of training, even nonprofessionals can program and control cobots. Cobots have raised high expectations of achieving flexible and resilient manufacturing processes, increasing the productivity, and assisting human workers [Makris 2021; Wang et al. 2019]. However, these expectations have not yet been met, resulting in a productivity gap [Schmidbauer et al. 2020b]. Weiss et al. [2021] provided an overview of the main research areas on human–robot interaction (HRI), work and organizational psychology, and sociology of work in the context of Industry 4.0. These are listed as follows:

- Safety and situation awareness (for example, safety certifications for cobot applications are costly and time-consuming [Rathmair and Brandstötter 2021]).
- Cobot programming and teaching (for example, cobot control and implementation expertise are still limited [Schmidbauer et al. 2020a]).
- Task dynamics, referring (for example) to the ironies of automation, stating that humans can no longer understand automated systems and the associated risks (for example, the fact that humans can no longer intervene when unforeseen errors occur [Bainbridge 1983]).

https://doi.org/10.34727/2022/isbn.978-3-85448-052-5_11

- Trust and acceptance to analyze (for example, factors facilitating or hindering trust and acceptance in HRI [Nordqvist and Lindblom 2018]).
- Skills, training, and workload such as the democratization of cobot technology [Hader et al. 2022].

One fundamental challenge when implementing a cobot on the shop floor is the HRI production planning; among others, the identification of suitable tasks and the determination of the best task allocation [Ranz et al. 2017] are considered. These issues must be overcome to unwrap the potential of HRI in a manufacturing environment. One approach to allocate tasks is the adaptive task sharing (ATS) between humans and cobots in a manufacturing environment [Schmidbauer 2022].

The main difference between ATS and conventional, static task allocation is that not only one best solution for a specific task allocation exists; a variety of options from which a human worker can choose is also available. An example of static task allocation is the optimization of a fitness function with respect to one criterion, usually time (minimum makespan) or (minimum) cost. In ATS, the workers are free to make their decisions. Other criteria, such as learning opportunities, task preferences, and physical and cognitive ergonomics can be considered. Therefore, this approach is not only suitable in terms of process flexibility but also focuses on a worker's well-being in a manufacturing environment.

In this chapter, state-of-the-art task allocation approaches and a main research gap in this area are presented. A new method for sharing tasks adaptively between humans and cobots in a manufacturing environment is proposed as a feasible solution. ATS is presented along with its three main pillars; a task analysis to identify suitable tasks for humans, cobots, and both (referred to as shareable tasks), a task assignment to preassign tasks to the agents or the shareable task set, and a task visualization for human workers to enable them to assign tasks from the shareable task set adaptively during the manufacturing process. The main benefits and the implications of this approach are presented and discussed.

2 Task Allocation Approaches

Task allocation between humans and machines is a massively discussed topic in manufacturing planning research. State-of-the-art scheduling algorithms capable of calculating the sequence and allocation of tasks to different agents, such as human workers, machines, and robots, have been proposed. A comprehensive elaboration of the state-of-the-art human–robot task allocation methods was reported by Schmidbauer [2022]. In this section, different approaches are exemplified.

2.1. Capability Indicator Evaluation

A compensatory approach to allocating tasks to humans or robots is to use capability or function indicators. The capabilities of the agents and the required capabilities of the tasks are described using quantitative or qualitative methods. This evaluation leads to a matching between the most suitable agent and a specific task.

An example procedure for capability-based task allocation was reported by Ranz et al. [2017]. Initially, the processes are categorized according to the process plan, and the process attributes are matched to the capabilities of the agents and the tasks. Subsequently, the invariable tasks are identified using a knock-out list and allocated to one of the agents. Next, the capability indicators for variable tasks are determined and described for both humans and robots. Then, the agents are comparatively evaluated using a pair-by-pair process. Apart from capability indicators, suitability indicators, such as ergonomic indicators, can be used [Mateus et al. 2019; Gualtieri et al. 2020]. Although this assignment appears to be static, in practice, it is not. The capabilities of humans can change by training, whereas their deskilling and physiological performance may change due to aging [Ranz et al. 2017]. The capabilities of robots can also change due to technological advances, wear and tear, and associated increased inaccuracies.

2.2. Fitness Functions

Based on a capability indicator evaluation or simply on the assumption that all tasks can be executed by both agents, a common task allocation approach is to set up a fitness or optimization function to maximize or minimize a target value. Target values are, for example, the operation time (makespan), cost, and throughput. An example was reported by Tsarouchi et al. [2017], where initially, the resources were evaluated in terms of their suitability and availability. Then, the resources with the lowest operation time that resulted in the minimum time were selected. In some approaches, several goals are also combined in one fitness function. For example, Pearce et al. [2018] focused on improving both the time and ergonomics and formulated them as a mixed-integer linear program.

2.3. Heuristics and Machine Learning

Heuristic approaches are used to provide a task allocation solution more efficiently than other approaches. The decision trees are presented, for example, by AND/OR [Darvish et al. 2018] or by Precedence [Riedelbauch and Henrich 2019] Graphs. If the decision trees are available, genetic algorithms can be used

to identify the best task allocation solution. Example deployments of genetic algorithms can be found in Beumelburg [2005]; Howard [2006]; Chen et al. [2014]; Weckenborg et al. [2020]. In those environments where not all decision cases are known, machine learning approaches can be employed. One example is the use of the Markov decision process framework, which is used to model a robot's actions [Roncone et al. 2017].

3 Research Gap

In this section, the task allocation research gap between humans and robots is examined. In operational research, capability indicator evaluations and optimization algorithms are employed to make the task sharing as effective or efficient as possible. In contrast, in the human factors/ergonomics (HF/E) research, a more decision-making authority for a human worker is proposed. Hacker and Sachse [2014] proposed higher decision authority and task diversity for workers to enable job enrichment and enlargement. Both forms of work organization aim at reduced monotony and less negative effects on humans. Additionally, in Ansari et al. [2018], more learning opportunities and less deskilling potential by employing higher task diversity were proposed.

Recent research work indicated that the workers' satisfaction can be increased through "ad hoc" task allocation [Tausch et al. 2020]. An online experiment (n = 151) indicated a higher level of satisfaction with the allocation process, the solution, and the result of the work process in the "ad hoc" scenario, where participants were able to allocate the tasks themselves. Therefore, the inclusion of workers in the task allocation process is crucial in exploiting the acceptance of human–robot interaction and in designing human-centered workplaces [Tausch and Kluge 2020].

Usually, the task allocation is implemented in the work-design-process phase (in industrial engineering) and is completed before the work begins. The reallocation of the so-called *shareable* tasks is enabled by monitoring workers and the work-system environment. Algorithms are being developed to make the robot adaptable to all situations. The active integration of human workers in the decision-making process was recommended by HF/E and engineering researchers [Buxbaum et al. 2020]; however, it was not implemented. The interests of both engineering and HF/E must be considered. These include, on the one hand, the economic efficiency of a process and, on the other hand, the improvement of human workers' ergonomics. For this purpose, the ATS method is developed as a method to share tasks adaptively between a human and a cobot in a manu-

facturing environment. The objective is to increase the economic efficiency and improve HF/E.

4 Adaptive Task Sharing

In this section, the ATS method is presented. The method was developed using an iterative design science research process based on Nunamaker Jr. et al. [1990]. The results of the five-stage research process contributed to the body of knowledge and vice versa [Schmidbauer 2022]. The proposed method consists of three parts with eight steps in total. In Figure 1, an overview of the ATS procedure is illustrated to show its different parts and steps. In the following subsections, the three parts of the proposed method are elaborated in more detail.

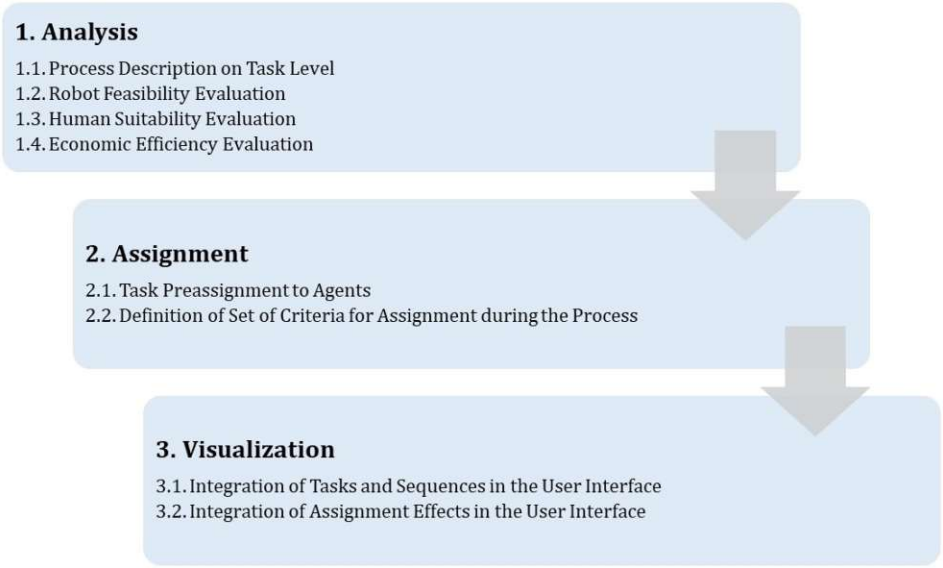


Figure 1 Adaptive Task-Sharing Method Procedure [Schmidbauer 2022].

4.1. Task Analysis

The task analysis of the proposed ATS method includes four steps. Initially, a task level process description is conducted. Therefore, a method based on several standards, such as DIN 8580, DIN 8593, and VDI 2860 is employed [Lotter 2012]. Then, the described tasks are evaluated regarding the automation feasibility using a cobot. The proposed method is based on a previous work [Gualtieri et al. 2020] and is further being developed. Five decision criteria are defined to identify if a task is feasible to be performed by a cobot. These criteria are spatial reach-

ability, payload, graspability, critical issues, and safety. It is assumed that a human worker can also execute all tasks. Therefore, no feasibility evaluation is required for a human worker. However, the tasks are evaluated for their suitability to a human worker. An ergonomics assessment using rapid upper-limb assessment (RULA) was reported in [McAtamney and Corlett 1993]. Finally, an economic efficiency evaluation of the proposed method is presented. Execution times and costs are assigned to each task and agent. The execution times are calculated by employing time stopping and the methods-time measurement (MTM). Then, the optimal time- and cost-efficient task allocations along with the optimal repetition rates are calculated.

4.2. Task Assignment

The key idea of ATS is to assign as many tasks to the shareable task set as possible. Therefore, only tasks that cannot be executed by the robot are permanently assigned to a human worker, and only tasks that are harmful (in terms of ergonomics) to the human worker are permanently assigned to the robot. This allows a high level of flexibility during the process. In the context of HF/E, three criteria for assignment, which are considered in ATS, are defined.

First, learning and training are important for a human worker. When workers are introduced to a new process, it is recommended that they take over the task, until they reach the tasks-pecific acceptance level of the learning curve [Jeske et al. 2014]. Second, task diversity affects a worker's satisfaction by reducing monotony [Hacker and Sachse 2014]. The perception of task diversity is not mathematically described because it is different for each individual. Therefore, ATS only incorporates the question "Does the task variety of the current task assignment correspond to my desired way of working?" in the user interface (UI). This question is a reminder to the workers that they can change the task assignment if they want. Third, the worker's preferences are considered to achieve job satisfaction. Research results on workers' preferences regarding tasks and allocations showed that workers tend to assign manual tasks to the robot and take over cognitive tasks such as checking tasks themselves [Schmidbauer 2022]. However, this is an individual study, and more data is needed to integrate workers' preferences into the ATS method. For this reason, preferences should not be suggested or calculated. However, if desired, they could be obtained from personal experience data. Considering the privacy of the workers, assignments could be collected, and later profiles of these workers could be created to suggest preferred task assignments they would probably like. At the moment, however, this is left solely to the worker to decide spontaneously and without applying any bias.

4.3. Task Visualization

To apply ATS, the visualization of tasks in a digital worker assistance system is necessary. An important requirement for visualization is that it can be easily understood by workers. To ensure high usability, a business process model and notation (BPMN)-based UI was selected. For each agent, i.e., human, robot, and *shareables* (human or robot), tasks can be modeled in lanes. The *shareables* tasks must be assigned to one of the execution agents, before the process starts. The interface features a start/stop button. Additionally, user instructions can be displayed on the interface. During the process, the current task is highlighted, so the user knows which task the cobot is executing and which tasks the user should execute. When the user finishes a task, they must confirm this by clicking on the task on the UI. The developed UI is depicted in Figure. 2.

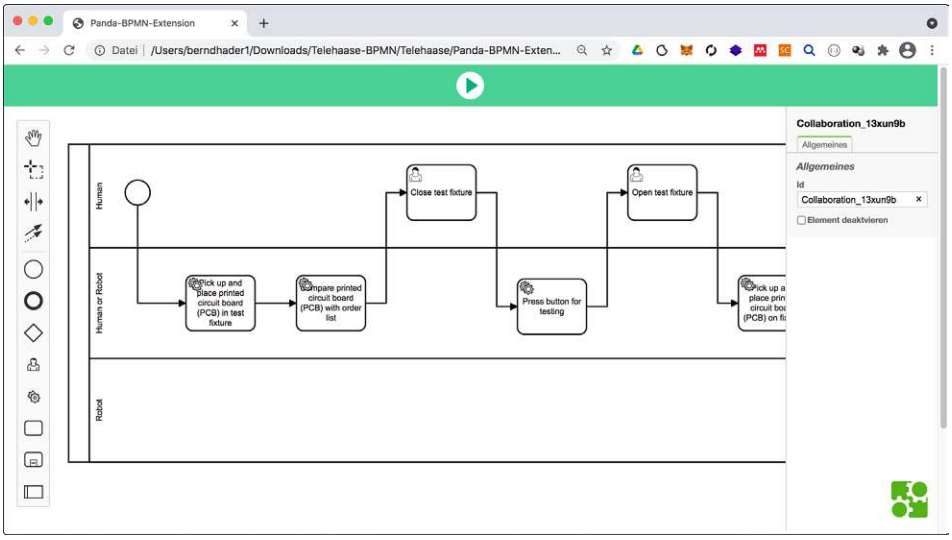


Figure 2 User interface visualizing the human, robot, and shareable (human or robot) tasks [Schmidbauer 2022].

The task visualization was realized using the BPMN-based Camunda[1] engine. The engine was connected with a *Franka Emika Panda* cobot to ensure efficient collaboration. The system architecture was introduced by Hader [2021] and Schmidbauer et al. [2021] and is available to the public on Github[2].

1 https://camunda.com/

2 https://github.com/berndhader/BPMN-Extension-Franka-Emika-Desk

5 Demonstration and Evaluation

In the following subsections, the demonstration and evaluation of the proposed ATS method are described.

5.1. Demonstration

The ATS method is demonstrated using two different case studies from the electronics industry (Figure 3). The first case study refers to the assembly of a heat sink, and the second study refers to the assembly of a timing relay. Both case studies are manual processes performed by electronics manufacturers in Vienna, but they differ in their number of tasks (case study I: 9 tasks; case study II: 18 tasks) and their task variety. In case study I, handling and joining assembly functions are mainly performed, whereas in case study II, some checking and special tasks (i.e.,., pressing a button or marking the order list) are performed. Both case studies were set up as hybrid workstations in the *Pilot Factory for Industry 4.0*[3] at TU Wien. A *Franka Emika Panda* cobot with a standard two-jaw gripper was used to execute the robot tasks.

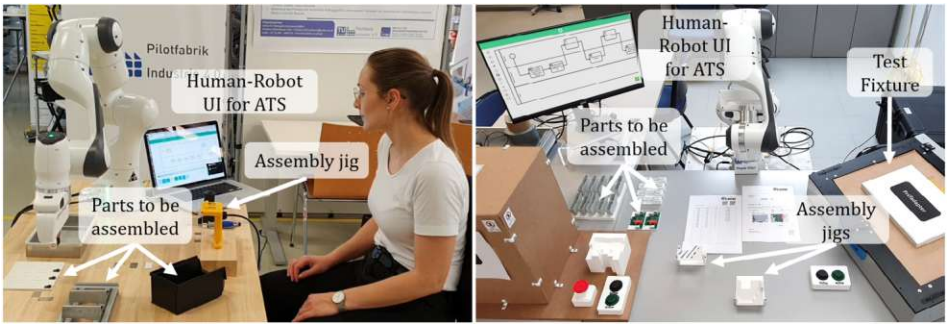


Figure 3 Case study I: “Assembly of a heat sink” demonstration experiment in *Pilot Factory for Industry 4.0* at TU Wien (adapted from Schmidbauer et al. [2020b]) and case study II: “Assembly of a timing relay” demonstration experiment at *TELE Haase Steuergeräte Ges.m.b.H* in Vienna, Austria (adapted from Schmidbauer [2022]).

Initially, the processes were defined at the task level, and a cobot feasibility evaluation was conducted to identify tasks that could not be assigned to a cobot because of issues related to spatial reachability, payload, graspability, safety, and other critical issues. Specifically, all tasks were evaluated according to these criteria and implemented on the cobot when possible. A human suitability evaluation was also conducted using RULA, where a simulation tool (Process Simulate

3 https://www.pilotfabrik.at/

Tecnomatix 15.0) was used to evaluate case study I. In case study II, RULA was applied using pen and paper.

Additionally, an economic efficiency evaluation was conducted using MTM-UAS for the manual tasks and by recording the execution times of the robot tasks. The optimal repetition rate and time- and cost-efficient task assignment variations were calculated. Detailed results of the task analysis have been presented in Schmidbauer [2022].

During task assignment, it was decided which tasks should be preassigned to an agent because not all tasks could be executed by both agents; for example, the RULA evaluation indicated that some tasks should be assigned to the cobot. An example task in case study I is “moving screws to transistors and putting together screws and transistors.” This task leads to a hand position, which is not ergonomic, and, therefore, it should always be taken over by the robot or an automated screwdriving machine. Both these processes were then modeled using BPMN.

5.2. Evaluation

5.2.1. Verification and Validation of the ATS Concept

The case studies were presented both for demonstrating the feasibility of the ATS method and for conducting different evaluations. Using the first demonstration experiment, economic efficiency calculations were performed, and the feasibility of the method was verified. Case study I was compared to other HRI cases related to manufacturing (i.e., the *cyber-physical production system (CPPS) Cell* and the *Potentiometer*) regarding different design aspects such as participatory design, scaling on demand, dynamic division of tasks, loose task coupling, reusable robot tasks, participatory robot programming, and overall development costs [Schmidbauer et al. 2020b]. The comparison showed that the ATS demonstration experiment scored well in participatory design, scaling on demand, dynamic division of tasks, and participatory robot programming. The overall development costs were relatively low. However, a specific laboratory setup was not ready to be directly integrated into the industry; the reason was that the loose task coupling and the reusable tasks had not been elaborated on time, since no UI for task reuse was implemented at the time. A comparison of different applications is presented in Figure 4.

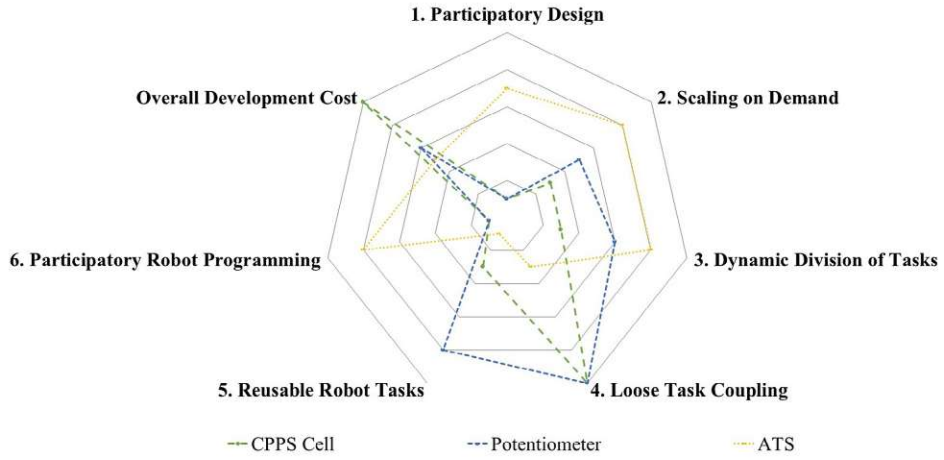


Figure 4 Comparison of different applications regarding the design aspects and overall development costs [Schmidbauer et al. 2020b].

5.2.2. Verification and Validation of the ATS User Interface

The first UI prototype was a mockup, which was used for a video vignette study [Zafari and Koeszegi 2020]. The mockup was used to design and develop the UI, which was evaluated within an online user study (n=51). During this study, the participants were introduced to the UI, and they modeled a human–robot process themselves. The usability, task load, task duration, and quality of the modeled tasks of the UI were evaluated. The usability was rated as excellent (System Usability Scale SUS: Ø = 86, SD = 12). A task load evaluation using the NASA raw-task load index also showed a very positive picture. The average results regarding the six task load variables were all below 1.6 on a 5-point Likert scale, where 1 indicates a very low and 5 indicates a very high demand, stress, effort, or frustration. The average perceived success was rated as 4.4. The average time spent by the participants accomplishing the BPMN modeling was 7:44 minutes (SD = 6:11). However, almost 16% of participants were not able to model the task without mistakes. This result indicates that at least a short period of training is necessary. The results showed that the BPMN processes could be understood by the participants, who were also able to model the tasks themselves. The UI can therefore be used by nonprofessionals with only a small period of training. The evaluation methodology and all results have been presented in Schmidbauer et al. [2021].

5.2.3. Final Validation of Concept and Verification of Requirements

The final evaluation of the method was performed for a user study (n=25) for case study II. This study was conducted in a factory, where, usually shop floor participants execute case study II. The main objective of this study was to validate the ATS concept in comparison to a static leftover task allocation and to explore a worker's preferences regarding task allocation and its effect on human factors. First, the participants attended a briefing and filled in an initial questionnaire. They were also introduced to 18 tasks and had to rank them according to their preference, i.e., whether a robot should take over the task or the participants wanted to execute the task themselves. Next, the participants went through two scenarios. In the ATS scenario, they were able to assign all *shareable* tasks to either the robot or them, whereas, in the other scenario, the tasks were already assigned to the robot, following a maximum automation approach. In each scenario, the participants also worked directly with the robot in the corresponding case study. After each scenario was completed, they filled in another questionnaire.

Most of the participants answered that they preferred the ATS scenario in comparison to the static task allocation (18/25, 72%). Additionally, the task allocation satisfaction was higher in the ATS scenario, and the participants reported that, in the production process, the task allocation should be assigned by humans and not by the robot or the "system". The participants' satisfaction with the task execution and the result was not significantly higher in the ATS scenario than in the static task allocation. The perceived competence and control were higher in the ATS scenario. The perceived (mental) task load was not higher in the ATS scenario, although the participants had additional decision tasks to do. These results show the positive impact of ATS on HF/E.

The ranking and assignment were analyzed regarding any pattern. The ranking exhibited no significance. The assignment showed that the participants assigned manual tasks more often to the robot than checking tasks. Significance in the assignment was found in four of the five handling tasks and in two of the eight other tasks (only 13 of the 18 tasks could be assigned by the participants). More results and information about the empirical user study have been presented in Schmidbauer [2022].

6 Discussion and Limitations

The results of the final evaluation of the method and the worker assistance system showed that participants prefer having the decision-making authority over task allocation. This result is in contrast to previous study results reported by Gombolay et al. [2015]. The outcome of the perceived satisfaction with the task

allocation conforms with the assumptions made by Tausch et al. [2020]. However, the results have not shown significant positive effects on the perceived satisfaction regarding task execution and results. The reasons for the difference in the results between the two studies can be attributed to the selection of participants or the performance of the robot during the experiment. The ATS evaluation results also showed that participants tend to assign tasks to the robot if they think that the robot is capable of performing these tasks. Wiese et al. [2021] obtained similar results. A common finding in all studies is that participants tend to assign more tasks to the robot than to them [Gombolay et al. 2015; Wiese et al. 2021; Tausch and Kluge 2020]. Another aspect is the increased perceived competence and control, which has implications for the intrinsic motivation and effectiveness of humans at work [Deci and Ryan 2000].

The practical implementation of ATS also exhibits some limitations. First, the additional engineering effort upfront must be mentioned. To implement ATS, the *shareable* tasks must be designed and implemented to be executable by both the robot and the human. This requires additional efforts in the design and implementation of workplaces and processes. If, for example, the task “screwing” is to be performed by both the cobot and the human worker, a manual screwdriver for the human and a screwdriving device for the cobot must be available [Schmidbauer et al. 2022].

Second, a safe interaction between the human worker and the cobot must be ensured. Cobots are considered as partly completed machinery, according to machinery directives (Directive 2006/42/EC of the European Parliament and of the Council of May 17, 2006 on machinery; amending Directive 95/16/EC (recast)). Therefore, standards regarding safety, such as the technical specification ISO/TS 15066:2016 on robots and robotic devices (specifically, collaborative robots) should be followed, and a risk assessment must be conducted. During a risk assessment, the entire workplace (including the cobot, the specific case study with its workpieces and fixtures, the robot program, and the required tools) must be considered. To date, these standards and risk assessments have considered workplaces that are set up once, and then, remain unchanged. Considering ATS, this means that all task sharing variants should be subjected to a separate risk assessment. Some approaches that could be integrated into a simulation have been reported [Vicentini et al. 2020]. Thus, the possibilities can already be evaluated in the digital twin [Bilberg and Malik 2019]. However, these possibilities are still immature for series production. Thus, they are considered as limitations in the ATS implementation.

7 Conclusion and Research Outlook

7.1. Conclusion

In this chapter, the design, development, demonstration, and evaluation of the proposed ATS method were described. ATS was proved to be an efficient method for adaptively sharing tasks between a human and a collaborative robot in a manufacturing environment. This method is capable of increasing the economic efficiency and improving human factors/ergonomics. The main differences between ATS and the static task allocation method are the postponement of the task allocation decision from the design phase to the shop floor and the ability of ATS to enable workers to have decision-making authority over task assignments. This is achieved via a digital worker assistance system, which visualizes the human–robot processes and serves as a UI to control the robot. The main benefits of this method are the following:

- Higher flexibility on the shop floor, which increases the economic efficiency, due to its higher potential to cope with mass customization requirements than the potential of other methods
- Cost savings via hybrid assembly, thus, increasing the economic efficiency
- Potential to reduce workers' physical stress through a task analysis, which improves human factors/ergonomics
- Increasing workers' satisfaction with "ad hoc" task allocation, which improves human factors/ergonomics.

7.2. Research Outlook

Adaptive task sharing between humans and collaborative robots enables dynamic and even individualizeable work organization in hybrid human–machine production systems. The implementation of ATS provides complementary task allocation to industrial practice and extends the possibilities for a flexible use of cobots in a manufacturing environment. ATS may be regarded as a further step toward democratization in terms of non-discriminating access for end users to the design, development, and use of cobot technology. To achieve this objective, complementary concepts, such as multimodal human–machine interfaces [Ionescu and Schlund 2021], intuitive teaching and programming concepts [El Zaatari et al. 2019], and dynamic simulations of adaptive work organization of human–robot teams [Pellegrinelli and Pedrocchi 2018] are needed. Furthermore, advances in (semi-)automated safety certification of reconfigurable human–cobot work systems as well as integrated safety and security concepts are required [Hollerer et

al. 2021]. Finally, the importance of workplace-based learning [Komenda et al. 2021] is crucial to maintain end users' competences and especially problem-solving skills within a more automated work environment, even in times when cobots will be widely-used as flexible and multipurpose manufacturing tools.

Bibliography

Alin Albu-Schäffer, Sami Haddadin, Christian Ott, Andreas Stemmer, Thomas Wimböck, and Gerd Hirzinger. 2007. The DLR lightweight robot: design and control concepts for robots in human environments. *Industrial Robot* 34, 5 (2007), 376–385. https://doi.org/10.1108/01439910710774386

Fazel Ansari, Philipp Hold, Walter Mayrhofer, Sebastian Schlund, and Wilfried Sihn. (2018). AUTODIDACT: Introducing the concept of mutual learning into a smart factory Industry 4.0. In *15th International Conference on Cognition and Exploratory Learning in Digital Age.*

Lisanne Bainbridge. 1983. Ironies of Automation. *Automatica* 19, 6 (1983), 775–779. https://doi.org/10.1016/0005-1098(83)90046-8

Katharina Beumelburg. 2005. *Fähigkeitsorientierte Montageablaufplanung in der direkten Mensch-Roboter-Kooperation.* Jost-Jetter Verlag, Fachverlag, 71296 Heimheim. https://doi.org/10.18419/opus-4037

Arne Bilberg and Ali Ahmad Malik. 2019. Digital twin driven human–robot collaborative assembly. *CIRP Annals* 68, 1 (2019), 499–502. https://doi.org/10.1016/j.cirp.2019.04.011

Hans-Jürgen Buxbaum, Sumona Sen, and Ruth Häusler. 2020. Theses on the future design of human-robot collaboration. In *Human-Computer Interaction. Multimodal and Natural Interaction,* Masaaki Kurosu (Ed.). Springer International Publishing, Cham, 560–579. https://doi.org/10.1007/978-3-030-49062-1_38

Fei Chen, Kosuke Sekiyama, Ferdinando Cannella, and Toshio Fukuda. 2014. Optimal subtask allocation for human and robot collaboration within hybrid assembly system. *IEEE Transactions on Automation Science and Engineering* 11, 4 (2014), 1065–1075. https://doi.org/10.1109/TASE.2013.2274099

Kourosh Darvish, Barbara Bruno, Enrico Simetti, Fulvio Mastrogiovanni, and Giuseppe Casalino. 2018. Interleaved online task planning, simulation, task allocation and motion control for flexible human-robot cooperation. In *2018 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. 58–65. https://doi.org/10.1109/ROMAN.2018.8525644

Edward L Deci and Richard M Ryan. 2000. The" what" and" why" of goal pursuits: Human needs and the selfdetermination of behavior. *Psychological inquiry* 11, 4 (2000), 227–268. https://doi.org/10.1207/ S15327965PLI1104_01

Shirine El Zaatari, Mohamed Marei, Weidong Li, and Zahid Usman. 2019. Cobot programming for collaborative industrial tasks: An overview. *Robotics and Autonomous Systems* 116 (2019), 162–180. https://doi.org/10.1016/j.robot.2019.03.003

Matthew C Gombolay, Reymundo A. Gutierrez, Shanelle G. Clarke, Giancarlo F. Sturla, and Julie A. Shah. 2015. Decisionmaking authority, team efficiency and human worker satisfaction in mixed human–robot teams. *Auton Robot* 39 (2015), 293–312. https://doi.org/10.1007/s10514-015-9457-9

Luca Gualtieri, Rafael A Rojas, Manuel A. Ruiz Garcia, Erwin Rauch, and Renato Vidoni. 2020. *Implementation of a laboratory case study for intuitive collaboration between man and machine in SME assembly.* Springer International Publishing, Cham, 335–382. https://doi.org/10.1007/978-3-030-25425-4_12

Winfried Hacker and Pierre Sachse. 2014. Allgemeine Arbeitspsychologie. Psychische Regulation von Tätigkeiten. *Zeitschrift für Arbeits- und Organisationspsychologie* A&O 58, 4 (2014), 221–222.

Bernd Hader. 2021. *Intuitive programming of collaborative human robot processes.* Master Thesis, TU Wien. https://doi.org/10.34726/hss.2021.76080

Bernd Hader, Christina Schmidbauer, Themistoklis Christakos, Eleni Tzavara, Sotiris Makris, and Sebastian Schlund. 2022. Democratizing industrial collaborative robot technology through interactive workshops in learning factories. *Proceedings of the 12th Conference on Learning Factoris (CLF 2022) (2022)*. https://doi.org/10.2139/ssrn.4074037

Siegfried Hollerer, Clara Fischer, Bernhard Brenner, Maximilian Papa, Sebastian Schlund, Wolfgang Kastner, Joachim Fabini, and Tanja Zseby. 2021. Cobot attack: a security assessment exemplified by a specific collaborative robot. *Procedia Manufacturing* 54 (2021), 191–196. https://doi.org/10.1016/j.promfg.2021.07.029

Ayanna M Howard. 2006. Role allocation in human-robot interaction schemes for mission scenario execution. In *Proceedings 2006 IEEE International Conference on Robotics and Automation, 2006. ICRA 2006.* 3588–3594. https://doi.org/10.1109/ROBOT.2006.1642250

Tudor B Ionescu and Sebastian Schlund. 2021. Programming cobots by voice: A human-centered, webbased approach. *Procedia CIRP 97* (2021), 123–129. https://doi.org/10.1016/j.procir.2020.05.213

Tim Jeske, Christopher M Schlick, and Susanne Mütze-Niewöhner. 2014. *Unterstützung von Lernprozessen bei Montageaufgaben.* Springer Berlin Heidelberg, Berlin, Heidelberg, 163–192. https://doi.org/10.1007/ 978-3-642-39896-4_4

Titanilla Komenda, Christina Schmidbauer, David Kames, and Sebastian Schlund. 2021. Learning to Share - Teaching the Impact of Flexible Task Allocation in Human-cobot Teams. *Proceedings of the Conference on Learning Factories (CLF) 2021*. http://dx.doi.org/10.2139/ssrn.3869551

Bruno Lotter. 2012. Einführung. In *Montage in der industriellen Produktion; Ein Handbuch für die Praxis,* Lotter B, Wiendahl H P. (eds). Springer Berlin Heidelberg, Berlin, Heidelberg, 1–8. https://doi.org/10.1007/978-3-642-29061-9_1

Sotiris Makris. 2021. *Cooperating robots for flexible manufacturing.* Springer Series in Advanced Manufacturing. https://doi.org/10.1007/978-3-030-51591-1

João Costa Mateus, Dieter Claeys, Veronique Limère, Johannes Cottyn, and El-Houssaine Aghezzaf. 2019. A structured methodology for the design of a humanrobot collaborative assembly workplace. *The International Journal of Advanced Manufacturing Technology* 102 (2019), 2663–2671. https://doi.org/10.1007/s00170-019-03356-3

Lynn McAtamney and E Nigel Corlett. 1993. RULA: a survey method for the investigation of world-related upper limb disorders. *Applied Ergonomics* 24, 2 (1993), 91–99. https://doi.org/10.1016/0003-6870(93)90080-s

Malin Nordqvist and Jessica Lindblom. 2018. Operators' experience of trust in manual assembly with a collaborative robot. In *Proceedings of the 6th International Con-*

ference on Human-Agent Interaction (Southampton, United Kingdom) (HAI '18). Association for Computing Machinery, New York, NY, USA, 341–343. https://doi.org/10.1145/3284432.3287180

Jay F Nunamaker Jr., Minder Chen, and Titus D M Purdin. 1990. Systems development in information systems research. *Journal of Management Information Systems* 7, 3 (1990), 89–106. https://doi.org/10.1080/07421222.1990.11517898

Margaret Pearce, Bilge Mutlu, Julie Shah, and Robert Radwin. 2018. Optimizing makespan and ergonomics in integrating collaborative robots into manufacturing processes. *IEEE Transactions on Automation Science and Engineering* 15, 4 (2018), 1772–1784. https://doi.org/10.1109/TASE.2018.2789820

Stefania Pellegrinelli and Nicola Pedrocchi. 2018. Estimation of robot execution time for close proximity human-robot collaboration. *Integrated Computer-Aided Engineering* 25, 1 (2018), 81–96. https://doi.org/10.3233/ICA-170558

Fabian Ranz, Vera Hummel, and Wilfried Sihn. 2017. Capability-based task allocation in human-robot collaboration. *Procedia Manufacturing* 9 (2017), 182–189. https://doi.org/10.1016/j.promfg.2017.04.011

Michael Rathmair and Mathias Brandstötter. 2021. Safety as bad cop of physical assistance systems?. In *Smart Technologies for Precision Assembly,* Svetan Ratchev (Ed.). Springer International Publishing, Cham, 344–357. https://doi.org/10.1007/978-3-030-72632-4_26

Dominik Riedelbauch and Dominik Henrich. 2019. Exploiting a human-aware world model for dynamic task allocation in flexible human-robot teams. In *2019 International Conference on Robotics and Automation (ICRA).* 6511–6517. https://doi.org/10.1109/ICRA.2019.8794288

Alessandro Roncone, Olivier Mangin, and Brian Scassellati. 2017. Transparent role assignment and task allocation in human robot collaboration. In *2017 IEEE International Conference on Robotics and Automation (ICRA).* 1014–1021. https://doi.org/10.1109/ICRA.2017.7989122

Christina Schmidbauer. 2022. *Adaptive task sharing between humans and cobots in assembly processes.* Dissertation, TU Wien. https://doi.org/10.34726/hss.2022.81342

Christina Schmidbauer, Bernd Hader, and Sebastian Schlund. 2021. Evaluation of a digital worker assistance system to enable adaptive task sharing between humans and cobots in manufacturing. In *54th CIRP Conference on Manufacturing Systems.* https://doi.org/10.1016/j.procir.2021.11.007

Christina Schmidbauer, Titanilla Komenda, and Sebastian Schlund. 2020a. Teaching cobots in learning factories – user and usability-driven implications. *Procedia Manufacturing* 45 (2020), 398–404. https://doi.org/10.1016/j.promfg.2020.04.043

Christina Schmidbauer, Hans Küffner-McCauley, Sebastian Schlund, Marcus Ophoven, and Christian Clemenz. 2022. *Detachable, low-cost tool holder for grippers in human-robot interaction. Springer Lecture Notes in Mechanical Engineering* (2022). forthcoming.

Christina Schmidbauer, Sebastian Schlund, Tudor B Ionescu, and Bernd Hader. 2020b. Adaptive task sharing in human-robot interaction in assembly. In *IEEE International Conference on Industrial Engineering and Engineering Management, IEEM 2020,* Singapore, December 14-17, 2020. IEEE, 546–550. https://doi.org/10.1109/IEEM45057.2020.9309971

Alina Tausch and Annette Kluge. 2020. The best task allocation process is to decide on one's own: effects of the allocation agent in human–robot interaction on perceived work characteristics and satisfaction. *Cognition, Technology & Work* (2020), 1–17. https://doi.org/10.1007/s10111-020-00656-7

Alina Tausch, Annette Kluge, and Lars Adolph. 2020. Psychological effects of the allocation process in human–robot interaction – A model for research on ad hoc task allocation. *Frontiers in Psychology* 11 (2020). https://doi.org/10.3389/fpsyg.2020.564672

Panagiota Tsarouchi, Alexandros-Stereos Matthaiakis, Sotiris Makris, and George Chryssolouris. 2017. On a human-robot collaboration in an assembly cell. *International Journal of Computer Integrated Manufacturing* 30, 6 (2017), 580–589. https://doi.org/10.1080/0951192X.2016.1187297

Federico Vicentini, Mehrnoosh Askarpour, Matteo G Rossi, and Dino Mandrioli. 2020. Safety assessment of collaborative robotics through automated formal verification. *IEEE Transactions on Robotics* 36, 1 (2020), 42–61. https://doi.org/10.1109/TRO.2019.2937471

Lihui Wang, Robert X Gao, József Váncza, Jörg Krüger, Xi Vincent Wang, Sotiris Makris, and George Chryssolouris. 2019. Symbiotic human-robot collaborative assembly. *CIRP Annals* 68, 2 (2019), 701–726. https://doi.org/10.1016/j.cirp.2019.05.002

Christian Weckenborg, Karsten Kieckhäfer, Christoph Müller, Martin Grunewald, and Thomas S Spengler. 2020. Balancing of assembly lines with collaborative robots. *Business Research* 13 (2020), 93–132. https://doi. org/10.1007/s40685-019-0101-y

Astrid Weiss, Ann-Kathrin Wortmeier, and Bettina Kubicek. 2021. Cobots in Industry 4.0: A roadmap for future practice studies on human-robot collaboration. *IEEE Transactions on Human-Machine Systems (2021).* https://doi.org/10.1109/THMS.2021.3092684

Eva Wiese, Patrick P Weis, Yochanan Bigman, and Kurt Gray. 2021. It's a match: Task assignment in human–robot collaboration depends on mind perception. *International Journal of Social Robotics* (2021). https://doi.org/10.1007/s12369-021-00771-z

Setareh Zafari and Sabine T Koeszegi. 2020. Attitudes toward attributed agency: Role of perceived control. *International Journal of Social Robotics* (2020), 1–10. https://doi.org/10.1007/s12369-020-00672-7

Trustworthy Robots in Society

Building trust in robots: A narrative approach

Jesse de Pagter

Abstract

This contribution proposes a narrative approach to trust-building with regards to robots. This should serve as a complementary interpretation in order to find new ways of theorizing and studying the trust-building process. The first aim of the paper is to distinguish between already existing notions of trust-building in relation to robots. I provide an overview arguing that with respect to building trust, robots are currently conceptualized as agentic interaction partners, as artifacts in sociotechnical systems that can be altered based on novel engineering and design processes, and finally as a type of technology that can potentially disrupt existing normative and legal conventions. From this overview, this paper proposes the complementary approach based on a narrative conceptualization of robots. This conceptualization focuses on the way that robots capture the imagination of many, arguing that this is fruitful to take into account when theorizing and studying the process of building trust in robots. The paper then discusses how this conceptualization can be developed in interdisciplinary research in the social sciences by evaluating and analyzing future imaginaries, developing anticipatory concepts, and facilitating access to sociotechnical potential.

Keywords

Anticipatory ethics, Robot ethics, Sociotechnical potential, Technological imaginaries, Technology narratives, Trust in robots

1 Introduction

Arguments that emphasize the need to build people's trust in robots have become increasingly prominent in recent years [European Commission 2019; Glikson and Woolley 2020; Ryan 2020]. A main reason for this perceived need to build trust is the expected increase in the use of robotic technologies across a wide array of domains. Examples of this range from the application of robots for personal assistance to self-driving cars and new types of robots in the workplace. Rapid advances are being made in technologies pivotal to this development, such as sensing technologies, machine vision, and machine learning. These technologies have granted robotic artifacts with increasing abilities to act autonomously and safely in real-world environments and expectations are that this trend will continue in the near future. This means that people's encounters with robots are likely to increase, as is public attention to the question of robots' impact on people's lives [Yang et al. 2018]. Given this context, it is not a surprise that the question of building trust in robots has attracted increased attention: lack of trust in an emerging technology like robots can have disruptive effects both on technological development itself and also on general trust in society [Frewer 1999].

The question of building trust in robots has become prominent in a wide variety of contexts. For instance, the need to build public trust in robotic technology has become more prominent due to the lessons learned regarding the societal impact of different emerging technologies in recent decades [Bunde et al. 2022;

https://doi.org/10.34727/2022/isbn.978-3-85448-052-5_12

Edelman 2020; Ethics Advisory Group 2018]. Further impetus can be found in the expected increase in interactions with robots, raising questions regarding how trust can be built in these interactions [Lewis et al. 2018; Naneva et al. 2020]. Furthermore, the growth of the use of robots and other forms of automation in the workplace may be a source of much fear of replacement and other forms of distrust toward robots [McClure 2018].

These examples indicate the importance of paying attention to the processes behind trust-building, but they also demonstrate that these processes can be interpreted and applied in different ways. Therefore, in the section that follows, this paper broadly distinguishes between different interpretations of the process of building trust in robots. Furthermore I also explain what this means in terms of the way robots themselves are conceptualized under these different interpretations. Building on this overview, Section 3 proposes a complementary notion that proposes a complementary approach that draws attention to the role of narratives. This approach considers robots to be a prominent example of a technology that is surrounded by many different narratives that often have imaginative and speculative content. I argue that this should be taken into account when theorizing and researching the process of building trust in robots. On this basis, Section 4 explains how such an approach can be developed in the social sciences. Finally, a short conclusion is presented to discuss how this notion of a trust-building process can be of use in the further development of interdisciplinary research.

2 Building trust in robots: Different interpretations

As noted, it is challenging to define the process of trust-building in a straightforward manner, amid the various interpretations of how it can be theorized and studied. To develop proper insight into the particularities of trust-building with regards to robots, a distinction among three interpretations of the trust-building process are developed below. It should be kept in mind that other ways of distinguishing these interpretations are possible; moreover, they often complement each other in actual research practice. Each of the following subsections indicates how a given interpretation theorizes the trust-building process and how the interpretation can be of use in research on trust in robots. In this way, the subsections provide the central ideas that define the different interpretations. Furthermore, this is accompanied by a description of the ways that robots themselves are defined and portrayed by this interpretation. As a consequence, this section does not stick to one single definition of the robot, but rather presents definitions of robots in relation to the respective approaches. Finally, the subsections investigate the research contexts and the domains in which these concepts and ideas are developed and deployed.

2.1. Behavior, appearance, and interaction

First, a set of prominent interpretations in this domain incorporate the idea of trust-building in relation to adjustments and refinements to the appearance and behavior of robots. Due to the emphasis on appearance and behavior, this perspective on trust usually aims to analyze human perceptions and experiences that are produced in interactions with robots [Hancock et al. 2011b; Li et al. 2010; Van den Brule et al. 2014]. For that reason, such research keeps a strong focus on gathering empirical insight from human-robot interaction experiments to identify and explain the mechanisms that support people's trust in robots during such interactions. Notions associated with this understanding of trust-building are often based on adjustments to the concept of interpersonal trust [Billings et al. 2012]. This concept involves the development of trust by one person (the trustor) in another (the trustee). Interpersonal trust has been studied extensively in fields like psychology and sociology, and it has been deployed in many different contexts. As such, the concept plays an important role in many theories of trust [Bachmann and Zaheer 2006; Simpson 2007]. Regarding the application of this notion of trust in relation to robots: in case one perceives technological artifacts as displaying forms of intelligence, they can also potentially enter into agentic relationships with humans [Elofson 2001; Nyholm 2018]. Hence, if artificial agency or intentionality emerges in interactive situations involving robots, notions derived from interpersonal trust can begin to play a role [De Graaf and Malle 2017].

Under this interpretation, robots are often defined as autonomous agents: (perceived) autonomous behavior and trust are thus seen as connected phenomena, as trust is generally considered to be an important element in relationships between humans as autonomous social beings. It is necessary to adjust this concept of trust to make it applicable to robots. In this setup, the robot takes on the role of the trustee in the interaction or relationship [Lewis et al. 2018]. In other words, although social agents are normally considered to be human, scholars in robotics-related research fields have argued that robots, when they are experienced and/or perceived by the trustor as an (intelligent) autonomous agent, can also be conceptualized as a trustee [Coeckelbergh 2012; Hancock et al. 2011a]. As such, robots fit into a wider discussion about trust in artificial agents—a discussion that also includes other types of agents, such as virtual bots, software programs, and so on [Andras et al. 2018; Glikson and Woolley 2020; Rossi 2018]. Nevertheless, it is crucial in this context that robots are embodied agents. The embodied character of robots opens up specific research areas that identify how trust can be built, based on the attitudes that this embodied appearance evokes in interactions [Nomura 2006]. For instance, the idea that robots have an embodied humanoid appearance is often considered to have a significant effect on the experience of trust [Alesich and Rigby 2017; van Pinxteren et al. 2019]. Thus, this

element of (anthropomorphic) embodiment makes appearance a very important feature and extends it to a wide variety of aspects that concern human beings' life and work with robots [Dumouchel 2022; Jones 2021].

It should come as no surprise that many of the concepts and methods that are based on this interpretation originate in psychological research. Approaches based on the notions of interpersonal trust and associated research into mental models have long been a topic of inquiry in several fields of psychology [Simpson 2007]. Human-robot interaction (HRI) is a prominent area of academic research that has successfully incorporated such concepts and methods to apply them to the study of trust in robots [Ullman and Malle 2018]. The interdisciplinary methods and approaches adopted in HRI generally focus on the development of experiments to measure trust-related attitudes. These experiments are often based on Wizard of Oz techniques, in which robots imitate agentic behavior [Riek 2012]. In this context, it is common to use validated questionnaires to gain insight into the experiences and attitudes of the human participants related to trust, while also providing directions on how trust can be built. Many outcomes of such research are then incorporated into the development of new robots, and robotics engineers often collaborate with HRI researchers in this context. Finally, several notions and theories developed in the context of interdisciplinary ethics research have also revolved around this interpretation of trust-building [Bartneck et al. 2021]. These notions and theories have been deployed to establish the field of robot ethics itself, but ethical concepts have also been implemented and tested as part of robots' behavioral cues [Malle 2016; Malle et al. 2019].

2.2. Research, development, and implementation

Another interpretation, focusing on the idea of human dependence on and vulnerability toward sociotechnical systems, describes the process of building trust in robots as an outcome of changes in design and engineering practices [Coeckelbergh 2013, 2015]. Taking technology to be constitutive of the environment in which humans operate and focusing on their vulnerability exposes trust as part of the entanglement that defines the relationship between humans and technological systems [Kiran and Verbeek 2010]. The implementation of this notion of trust-building draws attention to the ways in which research and innovation systems are set up, as well as the question of how they can be transformed in the direction of more open innovation in general [Geels 2004]. A good example of a framework often used in this context is the responsible research and innovation (RRI) framework [Asveld et al. 2017; van den Hoven et al. 2015]. When attention is drawn to the practices and norms that constitute sociotechnical systems, transparency and responsibility can become explicit components of (implicit) value

systems in research and engineering [Kiran et al. 2015]. For that reason, such approaches to trust-building focus on making innovation processes more open, responsible, and inclusive [Cheon and Su 2016]. Then, trust can emerge as an outcome of how characteristics such as reliance, transparency, and privacy are best managed in sociotechnical systems [Lee and See 2004; Wortham et al. 2016].

This interpretation of the trust-building process is not primarily focused on the appearance of the robotic artifact as such but rather emphasizes the notion of robotic technologies as important components of larger sociotechnical systems that (co-)define the conditions under which humans live and work [Sabanovic 2010]. Crucially, because humans construct these sociotechnical systems, they can also influence their development. Thus, it is important to consider how a technology like robotics establishes new forms of dependence and vulnerability, as well as the ways in which such issues are represented in terms of the norms and values of roboticists [Dignum et al. 2018]. Within the field of robotics itself, this perspective on the trust-building process has resulted in many calls to include norms and values that allow the needs of minorities to be recognized [Howard and Kennedy III 2020]. If design and engineering processes fail to consider and incorporate the values of different societal groups, attitudes of mistrust can arise with respect to technological systems [Howard and Borenstein 2018]. This, in turn, directly relates to overarching topics such as human rights and the maintenance of democratic values in technological design and engineering, emphasizing their importance for the way trust in robots develops in societies that have the need to mitigate the impacts of new types of robots [Torresen 2018]. An explicit openness to the deliberation on and implementation of values is in such a context considered to help ensure that societal and ethical issues are incorporated in the development processes behind robotic artifacts [Stahl and Coeckelbergh 2016]. Furthermore, the idea that robotic sociotechnical systems can establish new environments in which humans operate draws attention to the perspective of trust-building through the entanglements that constitute the relationship between humans and robotic systems [Richardson 2015]. As a central component of these sociotechnical systems, robotic artifacts can thus become more trustworthy by making their design and engineering to become more focused on issues like transparency and responsibility [Dignum 2017; Wortham et al. 2017; Wortham and Theodorou 2017].

Implicit in this idea of trust-building is the concept that existing practices in such fields can be altered to increase the general trustworthiness of robotic technologies. For instance, this can be achieved by implementing design requirements that would include the values discussed here and new types of awareness in robotics engineering and design processes [Siau and Wang 2018]. It is

important to note that many of these ideas are the subject of current discussions in the fields of robotics and HRI [Liu and Zawieska 2020; Winfield et al. 2021]. This is a crucial development, as their openness to such topics will likely have a strong effect on future developments in these pivotal fields. Critical analysis of and constructive engagement in new approaches to design and engineering practices are a prominent topic in many other academic areas as well. In philosophy, in particular, this entails the development of theories that reflect on technological design and engineering practices [Van de Poel and Royakkers 2011]. Several approaches from science and technology studies (STS) have also been crucial for drawing attention to the entanglements that constitute the (mundane) relationships between humans and technological artifacts [Maibaum et al. 2021; Rommetveit et al. 2020]. A range of topics and concepts from philosophy and the social sciences have likewise been used for interdisciplinary collaborations with roboticists, such as by creating approaches based on Participatory Design (PD) or Value Sensitive Design (VSD) [Azenkot et al. 2016; Umbrello and De Bellis 2018; Van Wynsberghe 2013].

2.3. Disruptions, rules, and regulations

The final interpretation regarding the process of building trust in robots is based on the idea that trust can be fostered with the help of rules and regulations [Nelson and Gorichanaz 2019]. Such discussions are increasingly prominent in recent years, as many proposals for rules and regulations to govern robotic and artificial intelligence (AI) technologies are currently in development [DG IPOL et al. 2016]. In close connection with this, the potential implications of the increasing prevalence of robots are a growing topic of inquiry in fields like ethics, legal studies, and governance studies [Boden et al. 2017; Leenes and Lucivero 2014; Nagenborg et al. 2008]. Beyond this, these types of interpretations of the trust-building process are generally important for the development of procedures that can help to mitigate the effects of emerging technologies on society. Ethical, legal, and regulatory schemes based on such analyses can help establish social trust in robots [Pagallo 2010]. Rules and regulations of this type can therefore function as part of a system of checks and balances that guide and govern technological developments and the implementation and use of robotics and AI, especially during a time characterized by socially disruptive technological advancements. In this context, philosophical deliberations are often concerned with new ontologies and ethical systems, while legal considerations are mostly about new rules and regulations. Both can be considered instrumental for creating a framework for further development and can help provide additional clarity for the current and future roles of robots in society [Gunkel 2012; Fosch-Villaronga and Heldeweg 2018].

In this context, robotic technologies (and AI) are largely understood and defined as a group of technologies set to disrupt existing conventions and therefore need to be guided and regulated via newly established rules and frameworks. Based on this idea, such approaches are often emphasizing the need for anticipating the potential social impact of future developments in robots' intelligence and agency. The emergence of intelligence and agency in machines is understood as a development that would potentially lead to large shifts in the issues of responsibility, liability, accountability and so on [Holder et al. 2016; Petit 2017]. If such issues are not dealt with properly, general trust in robots is likely to be compromised, which is why commitment to these issues can help to create rules and regulations to anticipate potential problems [Winfield and Jirotka 2018]. Thus, much of the work formulating ethical and/or legal arguments regarding the development of robots also takes on the current challenges and lacunae as well as those that future robots could bring about [Koops et al. 2013; Leenes et al. 2017]. In particular, with reference to concerns regarding the (im)possibilities of human control over the development and implementation of robotic technologies, ethical and legal scholars can help provide clarity to the discussion [Lin et al. 2012; Nagenborg et al. 2008].

When it comes to academic fields where trust-building of this type is a prominent topic, robot and AI ethics is a key area of research. The ethics of technology have been a topic of inquiry for many years, but it has gained importance in recent decades due to growing concerns over the social impact of other emerging technologies, such as nanotechnology or (big) data technologies [Brey 2017; Van de Poel 2008; Zwitter 2014]. In recent years, increasing interest has been seen in applying ethics to robots, and this has also become an important topic in fields investigating the governance of robotics. Hence, the meaning of the term ethics and its application have widened: according to some, ethics has even become "big business" [Richardson 2019; Sætra et al. 2021]. On a broader level, ethical considerations have repeatedly been shown to be instrumental for the exploration of potential legal and social ontologies and their consequences [Turner 2019]. In this regard, (social) robots are also becoming a subject of increasing concern in legal theory [Avila Negri 2021; Bertolini and Aiello 2018]. Furthermore, the regulation of robots and AI is now an important subject for concrete regulatory proposals, such as, for instance, in the European Union [European Parliament 2017].

3 Complementary interpretation: Robot narratives

In the previous section, different interpretations of the trust-building process were provided, accompanied by different definitions of robots: robots and robotic tech-

nologies were described as agentic interaction partners, as central artifacts in sociotechnical systems that are subject to alteration based on responsible engineering and design processes, and finally as a type of technology that (potentially) has the ability to disrupt existing ethical and legal conventions. I argue here for a complementary interpretation, describing a trust-building process that establishes the robot as the subject of narratives that may (and often do) contain speculative and imaginative content.

To explain this narrative perspective, it is useful to first discuss the notion of the narrative as found in social research, where it is used to analyze social life and has played that role for a long time [Nash 1994]. In social research, narratives are understood as carriers of meaning and assumptions, organized into plot-like structures [Deuten and Rip 2000]. Narratives constitute a crucial element of human social life: we think and communicate with the help of stories, which determine the limits of what we consider imaginable, knowable, and doable [Felt 2017]. In other words, narratives are instrumental for establishing meaning and structure [Czarniawska 2004]. As such, narratives can be analyzed in many different contexts, from policy documents to patient testimonies [Kirkpatrick 2008; McBeth and Lybecker 2018]. With regard to robots, the analysis of narratives can help clarify how robots become situated within shared meanings and assumptions. Thus, narratives are not simply stories: they can play a constitutive role in the development of concepts and ideas concerning the way our future with robots is to be configured. They point in certain directions, and the values implicit in them facilitate current and future development into a meaningful whole. Based on this, I argue that narratives can provide useful perspectives on the way we understand the role of robots, both in interactions with humans as well as in their larger societal context. Therefore, this paper argues for a more explicit inclusion of a narrative focus to come to grips with the way that the notion of trust-building can be further developed.

To ground the argument of the paper more securely, it will be useful to draw attention to narratives regarding robots and their imaginative and speculative elements. Why do narratives play such a crucial role for trust-building in robotics technology in particular? To provide a first answer to this question, it may be useful to provide insight into certain prominent elements from the history of robotics, as they demonstrate how the technological artifacts we call robots are surrounded by a host of speculative and imaginative narratives. The very term “robot” comes from a science fiction play, Rossum’s Universal Robots (R.U.R.), published in 1921 by the Czech writer Karel Čapek [Čapek 2004]. In this play, robots are created to work for humans, but they eventually rebel and cause the human race to go extinct. Even before this introduction of the word, autonomous non-human entities were a source of fear and fascination [Gasparetto 2016].

Depictions of and experiments with inanimate autonomous beings were part of larger (sometimes mesmerist and occultist) fascinations with automata. Such fascinations were rather widely expressed during the earlier phases of modern science and engineering [Coeckelbergh 2017; Liu 2010; Willis 2006]. The period of the Enlightenment for instance, exhibited an increasing engagement of clockmakers, mechanics at princely courts as well as other artisans with the creation of automata [Voskuhl 2013]. Furthermore, the history of fictional writing includes many examples of fascination with non-human forms of intelligence, such as the monster in Mary Shelley's Frankenstein, Henri Maillardet's Automaton, Nathanael (Nate) in E.T.A. Hoffmann's The Sandman, and many others [Cave and Dihal 2018; Selisker 2016].

In the context described above, as actual artifacts automata were mostly created in the domain of artisans, not that of engineers. With reference to the later establishment of robotics as a field of research and engineering, it is interesting to note that famous roboticists, such as Hans Moravec and Marvin Minsky, deliberately engaged in arguments that extrapolated research trends in their field toward futurist narratives. They claimed that science fiction futures that feature high levels of robot autonomy and intelligence could become a reality within a relatively short time. They explicitly referred to narratives that contained a strong fascination with the autonomy of robots. In that way, they were well aware that pop science efforts could help raise the political and cultural power of robotics as a field, which could in turn help increase their research funding [Geraci 2010].

In hindsight, it could be concluded that these early roboticists were quite successful in establishing robotics as a professional field. In this context, it is important to realize that the speculative dimension of the narratives around robots go well beyond the fictional realm. In recent decades, narratives about the further implementation of robots have continued to capture society's imagination [Heffernan 2019]. In the current moment in particular, there is a strong focus on the narrative that robots are an emerging technology that could, combined with AI technology, considerably alter the way we live and work while thoroughly changing society and the economy [Suchman 2019]. In this context, we have seen a general increase in concerns regarding the potential socially disruptive effects of the increasing implementation of autonomous systems, including robots, and their rapid technological progress. Important players like the European Union, Organization for Economic Co-operation and Development (OECD), and the United Nations have expressed the intention to maintain a strong emphasis on the need for anticipation of the future development of robotics in combination with AI technologies [European Commission 2020; OECD 2019; UNESCO 2021]. In this way, robots continue to be connected to the development of efforts to assess and predict future social and economic impact [Ford 2015; Nourbakhsh 2013]. There-

fore, due to its framing as an emerging technology, the future of robots is covered extensively in general public discourse, as well as in governance, which projects many different expectations onto its possible future development.

The speculative and anticipatory rhetoric that surrounds robotics is typical for emerging technologies, which are often characterized by high levels of ambiguity regarding their future [Asquer and Krachkovskaya 2021; Schaper-Rinkel 2013]. I argue here that the anticipation placed on (future) robots can usefully be understood and analyzed with the help of a narrative approach. As such, research and theory can treat narrative as a specific and distinct factor in the overall process of trust-building in robots. That is to say, robots' imagery and cultural status influence the way that they are portrayed and understood in the context of trust-building, in which individual robotic artifacts themselves, as well as robotics in general (as a field of research, design and implementation), play a crucial role in the emergence of narratives. Furthermore, I draw an explicit contrast to conceptions that disregard imaginative and speculative narratives about robotics as future-grasping hubris. Certainly, many solid and insightful studies exist that expose technological hubris and its distorting effects, but I argue that in relation to the process of building trust in robots, it can be insightful to explore how such narratives influence technological development and the culture that emerges around it. Expectation, imagination, and the anticipated/speculative future connected to them are thus considered major narratives that are constitutive of the ways that a culture thinks and acts with respect to robots.

4 Materializing a narrative approach: Studying trust

With a focus on narratives firmly established, it remains to describe how a narrative approach can be materialized. Here the interpretations from Section 2 are to be complemented by developing an understanding of how trust can be theorized and studied with the help of narratives. In other words, research that uses such an interpretation should be based on concepts of trust that explicitly, critically, and constructively engage with the narratives around robots. A particular focus is placed on three main components that are constitutive for a narrative approach to trust-building in robotic technologies: (1) scrutinizing existing imaginaries in the narratives about robots, (2) configuring anticipatory concepts regarding the narratives about robot futures, and (3) facilitating the emergence of new narratives around the sociotechnical potential of robots, mostly by increasing access to robots and robotics.

4.1. Scrutinizing technological imaginaries

To understand trust-building in robots using a narrative approach, it is crucial to draw attention to the technological imaginaries that are inherent to robot narratives. The concept of technological imaginaries or sociotechnical imaginaries emphasizes the entanglement of technologies in their social and cultural contexts [Jasanoff and Kim 2015]. The main idea being that these contexts define the development and implementation of technologies, as well as the norms and social and cultural practices around them. The analysis of technological imaginaries fit easily into a narrative approach, as these imaginaries can be found through the analysis of narratives. The main rationale here is that technological imaginaries drive cultural understandings and the perceptions of robots by defining and influencing arguments and concepts regarding robots' roles in our (future) societies. Thus, the imagined futures of robots should be understood as shaping the ways that societies deal with the contingencies connected to these futures through the visions and expectations that they represent. These imaginaries also shape the development of technologies connected to such visions [Jasanoff and Kim 2009]. As such, the technological imaginary should also play a constitutive role in the critical analysis of anticipatory notions surrounding robots in relation to the construction of novel social realities based on the futures of emerging technologies [Vallès-Peris and Doménech 2020]. For instance, Lucy Suchman has convincingly argued that the robot imaginary confronted at present is largely based on Euro-American notions of embodiment, emotion, and sociality. From this argument, she demonstrates that narratives of social order are reproduced in the specific technological designs of robots [Suchman 2006]. Another example is the book The Robotic Imaginary by Jennifer Rhee, which analyzes the conceptualizations and visions of humanness and dehumanization as seen in discourses on robotics [Rhee 2018].

The analysis of imaginaries is particularly useful when one wants to study and analyze different interpretations and controversies in narratives that are concerned with the (future) role of robots in our societies. Many other technologies and their particular imaginaries have already undergone scrutiny using analysis of this type [Jasanoff and Kim 2015; Sismondo 2020]. These studies have repeatedly demonstrated that perceptions of technologies and their futures are a crucial factor in the decision-making of governments and corporations. Furthermore, they are instrumental to the development and negotiation of novel and already present social arrangements in terms of new technologies, for instance in the context of governance [Grunwald 2018]. In relation to the process of building trust in robots in light of promises, expectations, and fears regarding robots and their futures, trust-building can be conceptually connected to the ways in which robots are presented in (speculative) narratives [Rommetveit and Wynne 2017]. In other

words, such narratives must be interpreted as drivers of debates regarding the possibilities, dangers, and challenges of robotization and automation.

Thus, narratives and their imaginaries are drivers of the establishment of social and public trust with respect to robots [Kearnes et al. 2006]. Furthermore, when used in conjunction with concepts of trust that are derived from interpersonal trust, they can help provide a deeper understanding of people's attitudes in human-robot interaction [Fortunati et al. 2015; Weiss and Spiel 2021]. In this way, concepts of trust in robots can be further refined through careful investment in inclusive and responsible imaginaries with respect to our future with robots. Thus, analyzing narratives that establish certain social imaginaries, the heavily anticipated roles of robots in society can be assessed, discussed, and criticized. Finally, the analysis of imaginaries of robotics in different domains (e.g., robot engineering, robot governance, and industrial contexts) can help establish new understandings of social and collective life with robots while recognizing the social character of such technological futures.

4.2. Configuring anticipatory concepts

In addition to the critical analysis of robot imaginaries, a second component involves actively taking part in the development and configuration of concepts that can support narratives that are engaged with the anticipation of robots the sociotechnical systems that emerge around them [Floridi 2014]. Here, philosophers and social scientists themselves can become involved in the anticipation of potential scenarios in order to develop the arguments and concepts that can be of use in the responsible implementation of robotic technologies [Brey 2012]. In comparison to the subsection above, this component also requires a critical stance toward robot futures, but simultaneously it is more strongly focused on constructive and sometimes speculative engagement with the futures of emerging robotics. The different ways in which the technological potential of robots is imagined can be assessed and refined to shape the sociotechnical systems that surround robots [Plas et al. 2010]. Although many types of robots that are anticipated are not yet in widespread use, speculative engagement with their future incarnations can be an important part of concepts of trust-building that are based on a narrative approach. The provision of new directions and concepts to guide the construction of narratives about our futures with robots can allow new roles to be allotted to them, ones that can already be anticipated [Gunkel 2022; Selkirk et al. 2018].

In general, the advancement of such anticipatory concepts can encourage reflection on notions of trust to address current challenges surrounding automation and robotics. The main emphasis should fall on creating concepts to help soci-

eties adjust in times of transformative change, times in which technological developments challenge and redefine societal norms and practices [Bratton 2017; Sardar 2010]. As such, the configuration of anticipatory concepts involves the production of a thorough overview of the meanings and interpretations that develop in the anticipated trajectories of robotic development, including its speculative elements. The aim is thus to create new concepts that can help anticipate and modify the sociotechnical ramifications of those developments [Castañeda and Suchman 2014]. Apart from analyzing technological products and innovations in their social context, the goal should be to engage in the development of new narratives that can help steer the development of future products and innovations.

I argue here for explicit commitment to the continuous (re)configuration of anticipatory concepts related to robots. This is largely an exploratory endeavor, in which it is crucial to invest in concepts that support more inclusive narratives of robots as a widely implemented technology [Grunwald 2010; Selin 2008]. Interdisciplinary work is crucial for such efforts and for developing concepts that are on the one hand speculative, but rooted in engineering reality on the other. Moreover, it is a significant platform for implementing concepts and ideas that mobilize the technoscientific imagination toward emancipatory sociotechnical systems. Thus, by facilitating novel definitions of robots and their roles in social contexts, anticipation based on speculative concepts would be instrumental to fostering novel engagement types with robots. In this way, robots can help change well-established social ontologies [Coeckelbergh 2010; Gunkel 2018].

4.3. Facilitating sociotechnical potential

Finally, within a narrative approach that is focused on trust-building, it is important to provide insights and pathways that can actively facilitate the emergence of new narratives about robots' sociotechnical potential. These narratives may be instrumental for developing ideas for the use of robots, founded on the imaginative capacity of the general public or of specific future users regarding how they conceptualize and imagine life and work with robots. Facilitating narratives around robotics' potential is therefore mostly about deliberately providing access to robots in order to allow new narratives to emerge [Chun 2019; Fischer et al. 2020]. Furthermore, this calls for critical but constructive engagements with people's concrete imaginings with respect to their use of and work with robots. An important idea in this approach is that technological artifacts such as robots are (re)defined in terms of how their use is imagined and practiced [Soljacic et al. 2022]. Thus, the identification of new forms of sociotechnical potential can enable the development of a way to allow for new understandings of the roles that robots can play in society.

Here, an important question is how the different uses of technology can be facilitated and analyzed [Cressman 2019]. The researcher's role in this process is to work to provide access to robots and connect the narrative understanding of technology to people's experiences while using and interacting with robots. Thus, it is helpful to facilitate the emergence of new narratives around the possible uses of robots for building social trust in robots in a democratic society. To create trust and implement technology in accordance with democratic values, the general public as well as individual users must be prompted to form new narratives around robots' future potential [Bijker 2010; Ionescu and Schlund 2019]. In line with this, constructive engagement with narratives that involve robots' sociotechnical potential can be developed by increasing interactions with robots and robotics. The goal of this activity is not necessarily to see how different groups and guidelines can be included in the design but rather to inquire into the ways in which people use and understand technologies in novel ways that are previously unimagined.

It is crucial to recall that this approach must be explicitly neutral to any narrative trajectory, even with respect to those trajectories that could be classified as irrationally utopian or dystopian. The goal is rather to facilitate the way that associations of this or other types lead to unanticipated mundane uses of robots. Pioneering studies in the social construction of technology have been undertaken in relation to the user as an agent of technological change [Kline and Pinch 1996]. These studies indicate the way that a certain technological artifact and its social environment evolve over time, based on actual use. Therefore, in relation to robots and building trust in them, research activities should not only critically analyze and anticipate robot futures but also focus on providing the possibilities for emancipation and democratization through imagination in narratives regarding the use of and work with robotics. In this way, the development and implementation trajectories of robots can become increasingly democratized through the emphasis on possibilities for choosing and designing different technologies [Feenberg 2002].

5 Conclusion

This paper presented an approach to the process of building trust in robots that focuses on the role that narratives can play. I have demonstrated that robotics is necessarily embedded in narratives about its own future. I argue that this necessitates a complementary view on building trust in robots, which I presented in this paper. The goal of this approach is to deploy already existing discourse on trust to generate new ideas for bringing robots into our societies in ways that, without

profoundly disturbing our economic, social, and political lives, might empower us to achieve more equal, sustainable, and desirable futures with robots.

Implicitly, this focus on narratives involves the addition of perspectives from history, the arts, literature, and philosophy to the already rapidly growing body of research on the implications of emerging technologies such as robots. This development is far from finished and certainly is not limited to roboticists adapting or being open to these kinds of perspectives. It also means that significant efforts must still be made to bring the above-mentioned fields and disciplines closer to the field of robotics and identify ways in which the interpretations and ideas of each can be of use for the other. This is and will continue to be very challenging, not least because interdisciplinary work often necessarily encounters and must deal with long-standing preconceptions and conflicting epistemologies between disciplines [Weszkalnys and Barry 2013]. Therefore, it is crucial to continue investing in efforts to produce a deeper integration of these inter- and transdisciplinary perspectives.

Bibliography

Simone Alesich and Michael Rigby. 2017. Gendered Robots: Implications for Our Humanoid Future. *IEEE Technology and Society Magazine* 36, 2 (June 2017), 50–59. https://doi.org/10.1109/MTS.2017.2696598

Peter Andras, Lukas Esterle, Michael Guckert, The Anh Han, Peter R Lewis, Kristina Milanovic, Terry Payne, Cedric Perret, Jeremy Pitt, Simon T Powers, Neil Urquhart, and Simon Wells. 2018. Trusting Intelligent Machines: Deepening Trust Within Socio-Technical Systems. *IEEE Technology and Society Magazine* 37, 4 (December 2018), 76–83. DOI:https://doi.org/10.1109/MTS.2018.2876107

Alberto Asquer and Inna Krachkovskaya. 2021. Uncertainty, institutions and regulatory responses to emerging technologies: CRISPR Gene editing in the US and the EU (2012–2019). *Regulation & Governance* 15, 4 (2021), 1111–1127. DOI:https://doi.org/10.1111/rego.12335

Lotte Asveld, Rietje van Dam-Mieras, Tsjalling Swierstra, Saskia Lavrijssen, Kees Linse and Jeroen van den Hoven (eds.). 2017. *Responsible innovation 3: a European agenda?* Springer. https://doi.org/10.1007/978-3-319-64834-7

Sergio M C Avila Negri. 2021. Robot as Legal Person: Electronic Personhood in Robotics and Artificial Intelligence. *Frontiers in Robotics and AI* 8, (2021). https://doi.org/10.3389/frobt.2021.789327

Shiri Azenkot, Catherine Feng, and Maya Cakmak. 2016. Enabling building service robots to guide blind people a participatory design approach. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 3–10. https://doi.org/10.1109/HRI.2016.7451727

Reinhard Bachmann and Akbar Zaheer (Eds.). 2006. *Handbook of trust research.* Edward Elgar Publishing, Cheltenham.

Christoph Bartneck, Christoph Lütge, Alan Wagner, and Sean Welsh. 2021. *An Introduction to Ethics in Robotics and AI.* Springer International Publishing, Cham. https://doi.org/10.1007/978-3-030-51110-4

Andrea Bertolini and Giuseppe Aiello. 2018. Robot companions: A legal and ethical analysis. *The Information Society 34*, 3 (May 2018), 130–140. https://doi.org/10.1080/01972243.2018.1444249

Wiebe E Bijker. 2010. Democratization of Technological Culture. In *Science and Technology Studies at Maastricht University. An Anthology of Inaugural Lectures,* Karin Bijsterveld (ed.). Maastricht University Press, Maastricht, 13–41.

Deborah Billings, Kristin Schaefer, Jessie Chen, and Peter Hancock. 2012. Human-robot interaction: Developing trust in robots. (March 2012). In *HRI '12: Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction.* https://doi.org/10.1145/2157689.2157709

Margaret Boden, Joanna Bryson, Darwin Caldwell, Kerstin Dautenhahn, Lilian Edwards, Sarah Kember, Paul Newman, Vivienne Parry, Geoff Pegman, Tom Rodden, Tom Sorrell, Mick Wallis, Blay Whitby, and Alan Winfield. 2017. Principles of robotics: regulating robots in the real world. *Connection Science* 29, 2 (April 2017), 124–129. https://doi.org/10.1080/09540091.2016.1271400

Benjamin H. Bratton. 2017. The New Normal. Strelka Press, Moscow and London.

Philip Brey. 2017. Ethics of emerging technology. In *The ethics of technology: Methods and approaches,* Sven Ove Hansson (ed.). Rowman & Littlefield International, 175–191. Philip Brey. 2017. Ethics of emerging technology. In The ethics of technology: Methods and approaches. 175–191.

Philip Brey. 2012. Anticipatory Ethics for Emerging Technologies. *Nanoethics* 6, 1 (April 2012), 1–13. https://doi.org/10.1007/s11569-012-0141-7

Rik van den Brule, Ron Dotsch, Gijsbert Bijlstra, Daniel H J Wigboldus, and Pim Haselager. 2014. Do Robot Performance and Behavioral Style affect Human Trust? *International Journal of Social Robotics* 6, 4 (November 2014), 519–531. https://doi.org/10.1007/s12369-014-0231-5

Tobias Bunde, Sophie Eisentraut, Natalie Knapp, Randolf Carr, Julia Hammelehle, Isabell Kump, Luca Miehe, and Amadée Mudie-Mantz. 2022. *Munich Security Report 2022: Turning the Tide – Unlearning Helplessness.* Munich Security Conference, Munich. Retrieved from https://doi.org/10.47342/QAWU4724

Karel Čapek. 2004. R.U.R (Rossum's universal robots). Penguin.

Claudia Castañeda and Lucy Suchman. 2014. Robot visions. *Social Studies of Science* 44, 3 (June 2014), 315–341. https://doi.org/10.1177/0306312713511868

Stephen Cave and Kanta Dihal. 2018. Ancient dreams of intelligent machines: 3,000 years of robots. *Nature* 559, 7715 (July 2018), 473–475. https://doi.org/10.1038/d41586-018-05773-y

EunJeong Cheon and Norman Makoto Su. 2016. Integrating Roboticist Values into a Value Sensitive Design Framework for Humanoid Robots. In *The Eleventh ACM/IEEE International Conference on Human Robot Interaction (HRI '16)*, IEEE Press, Christchurch, New Zealand, 375–382. doi: 10.1109/HRI.2016.7451775

Bohkyung Chun. 2019. Doing autoethnography of social robots: Ethnographic reflexivity in HRI. Paladyn, *Journal of Behavioral Robotics* 10, 1 (January 2019), 228–236. https://doi.org/10.1515/pjbr-2019-0019

Mark Coeckelbergh. 2010. Robot rights? Towards a social-relational justification of moral consideration. Ethics and Information Technology 12, 3 (September 2010), 209–221. https://doi.org/10.1007/s10676-010-9235-5

Mark Coeckelbergh. 2012. Can we trust robots? *Ethics and Information Technology* 14, 1 (March 2012), 53–60. https://doi.org/10.1007/s10676-011-9279-1

Mark Coeckelbergh. 2013. *Human being @ risk: enhancement, technology, and the evaluation of vulnerability transformations.* Springer, Dodrecht. https://doi.org/10.1007/978-94-007-6025-7

Mark Coeckelbergh. 2015. The tragedy of the master: automation, vulnerability, and distance. *Ethics and Information Technology* 17, 3 (September 2015), 219–229. https://doi.org/10.1007/s10676-015-9377-6

Mark Coeckelbergh. 2017. *New Romantic Cyborgs. Romanticism, Information Technology, and the End of the Machine.* MIT Press, Cambridge, MA.

Darryl Cressman. 2019. Disruptive Innovation and the Idea of Technology. *Novation: Critical Studies of Innovation* 1 (June 2019), 17–39.

Barbara Czarniawska. 2004. *Narratives in social science research.* Sage Publications, London; Thousand Oaks, Calif.

Jasper Deuten and Arie Rip. 2000. Narrative Infrastructure in Product Creation Processes. *Organization* 7, 1 (February 2000), 69–93. https://doi.org/10.1177/135050840071005

DG IPOL – Directorate-General for Internal Policies of the Union, European Parliament, Nathalie Nevejans. 2016. *European civil law rules in robotics.* Publications Office. Retrieved January 13, 2020 from https://data.europa.eu/doi/10.2861/946158

Virginia Dignum. 2017. Responsible Artificial Intelligence: Designing Ai for Human Values. *ITU Journal: ICT Discoveries* 1 (2017), 9.

Virginia Dignum, Frank Dignum, Javier Vazquez-Salceda, Aurelie Clodic, Manuel Gentile, Samuel Mascarenhas, and Agnese Augello. 2018. Design for Values for Social Robot Architectures. *Envisioning Robots in Society – Power, Politics, and Public Space* (2018), 43–52. https://doi.org/10.3233/978-1-61499-931-7-43

Paul Dumouchel. 2022. Ethics & Robotics, Embodiment and Vulnerability. *International Journal of Social Robotics* (March 2022). https://doi.org/10.1007/s12369-022-00869-y

Edelman. 2020. *Edelman Trust Barometer 2020 special Report: Trust in Technology.* Edelman. Retrieved May 27, 2020 from https://www.edelman.com/sites/g/files/aatuss191/files/2020-02/2020%20Edelman%20Trust%20Barometer%20Tech%20Sector%20Report_1.pdf

Greg Elofson. 2001. Developing Trust with Intelligent Agents: An Exploratory Study. In *Trust and Deception in Virtual Societies*, Cristiano Castelfranchi and Yao-Hua Tan (eds.). Springer Netherlands, Dordrecht, 125–138. https://doi.org/10.1007/978-94-017-3614-5_6

Ethics Advisory Group. 2018. T*owards a digital ethics.* EDPS. Retrieved July 23, 2019 from https://edps.europa.eu/sites/edp/files/publication/18-01-25_eag_report_en.pdf

European Commission. 2019. *Building Trust in Human-Centric Artificial Intelligence.* Retrieved from https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52019DC0168&qid=1630917561643

European Commission. 2020. White Paper On Artificial Intelligence - A European approach to excellence and trust. Retrieved November 11, 2020 from https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM:2020:0065:FIN

European Parliament. 2017. European Parliament resolution of 16 February 2017 with recommendations to the Commission on Civil Law Rules on Robotics.

Andrew Feenberg. 2002. *Transforming technology: a critical theory revisited.* Oxford University Press, New York, N.Y.

Ulrike Felt. 2017. "Response-able Practices" or "New Bureaucracies of Virtue": The Challenges of Making RRI Work in Academic Environments. In *Responsible Innovation 3: A European Agenda?*, Lotte Asveld, Rietje van Dam-Mieras, Tsjalling Swierstra, Saskia Lavrijssen, Kees Linse and Jeroen van den Hoven (eds.). Springer International Publishing, Cham, 49–68. https://doi.org/10.1007/978-3-319-64834-7_4

Kerstin Fischer, Johanna Seibt, Raffaele Rodogno, Maike Kirkegård Rasmussen, Astrid Weiss, Leon Bodenhagen, William Kristian Juel, and Norbert Krüger. 2020. Integrative Social Robotics Hands-on. *Interaction Studies* 21, 1 (January 2020), 145–185. https://doi.org/10.1075/is.18058.fis

Luciano Floridi. 2014. Technoscience and Ethics Foresight. Philosophy & Technology 27, 4 (December 2014), 499–501. https://doi.org/10.1007/s13347-014-0180-9

Martin Ford. 2015. *Rise of the robots: technology and the threat of a jobless future.* Basic Books, New York.

Leopoldina Fortunati, Anna Esposito, Mauro Sarrica, and Giovanni Ferrin. 2015. Children's Knowledge and Imaginary About Robots. *International Journal of Social Robotics* 7, 5 (November 2015), 685–695. https://doi.org/10.1007/s12369-015-0316-9

Eduard Fosch-Villaronga and Michiel Heldeweg. 2018. "Regulation, I presume?" said the robot – Towards an iterative regulatory process for robot governance. *Computer Law & Security Review* 34, 6 (December 2018), 1258–1277. https://doi.org/10.1016/j.clsr.2018.09.001

Lynn Frewer. 1999. Risk Perception, Social Trust, and Public Participation in Strategic Decision Making: Implications for Emerging Technologies. *Ambio* 28, 6 (1999), 569–574.

Alessandro Gasparetto. 2016. Robots in History: Legends and Prototypes from Ancient Times to the Industrial Revolution. In *Explorations in the History of Machines and Mechanisms (History of Mechanism and Machine Science),* López-Cajún, C., Ceccarelli, M. (eds). Springer International Publishing, Cham, 39–49. https://doi.org/10.1007/978-3-319-31184-5_5

Frank W Geels. 2004. From sectoral systems of innovation to socio-technical systems: Insights about dynamics and change from sociology and institutional theory. *Research Policy* 33, 6 (September 2004), 897–920. https://doi.org/10.1016/j.respol.2004.01.015

Robert M Geraci. 2010. *Apocalyptic AI: visions of heaven in robotics, artificial intelligence, and virtual reality.* Oxford University Press, New York.

Ella Glikson and Anita Williams Woolley. 2020. Human Trust in Artificial Intelligence: Review of Empirical Research. Academy of Management Annals 14, 2 (March 2020), 627–660. https://doi.org/10.5465/annals.2018.0057

Maartje de Graaf and Bertram F Malle. 2017. How People Explain Action (and Autonomous Intelligent Systems Should Too). *AAAI Fall Symposia.* (2017), 8.

Armin Grunwald. 2010. From Speculative Nanoethics to Explorative Philosophy of Nanotechnology. *Nanoethics* 4, 2 (August 2010), 91–101. https://doi.org/10.1007/s11569-010-0088-5

Armin Grunwald. 2018. *Technology Assessment in Practice and Theory.* Routledge, London. https://doi.org/10.4324/9780429442643

David Gunkel. 2012. *The machine question: critical perspectives on AI, robots, and ethics.* MIT Press, Cambridge, Mass.

David Gunkel. 2018. The other question: can and should robots have rights? *Ethics and Information Technology* 20, 2 (June 2018), 87–99. https://doi.org/10.1007/s10676-017-9442-4

David Gunkel. 2022. The Symptom of Ethics: Rethinking Ethics in the Face of the Machine. *Human-Machine Communication* 4, 1 (April 2022). https://doi.org/10.30658/hmc.4.4

Peter A Hancock, Deborah R Billings, Kristin E Schaefer. 2011a. Can You Trust Your Robot? *Ergonomics in Design* 19, 3 (July 2011), 24–29. https://doi.org/10.1177/1064804611415045

Peter A Hancock, Deborah R Billings, Kristin E Schaefer, Jessie Y C Chen, Ewart J de Visser, and Raja Parasuraman. 2011b. A Meta-Analysis of Factors Affecting Trust in Human-Robot Interaction. *Human Factors* 53, 5 (October 2011), 517–527. https://doi.org/10.1177/0018720811417254

Teresa Heffernan. 2019. Fiction Meets Science: Ex Machina, Artificial Intelligence, and the Robotics Industry. In *Cyborg Futures*, Teresa Heffernan (ed.). Springer, Berlin, 127–140.

Chris Holder, Vikram Khurana, Faye Harrison, and Louisa Jacobs. 2016. Robotics and law: Key legal and regulatory implications of the robotics age (Part I of II). *Computer Law & Security Review* 32, 3 (June 2016), 383–402. https://doi.org/10.1016/j.clsr.2016.03.001

Jeroen van den Hoven, Pieter E Vermaas, and Ibo van de Poel (Eds.). 2015. *Handbook of Ethics, Values, and Technological Design.* Springer Netherlands, Dordrecht. https://doi.org/10.1007/978-94-007-6970-0

Ayanna Howard and Jason Borenstein. 2018. The Ugly Truth About Ourselves and Our Robot Creations: The Problem of Bias and Social Inequity. *Science and Engineering Ethics* 24, 5 (October 2018), 1521–1536. https://doi.org/10.1007/s11948-017-9975-2

Ayanna Howard and Monroe Kennedy III. 2020. Robots are not immune to bias and injustice. *Science Robotics* 48, 5 (2020). https://doi.org/10.1126/scirobotics.abf1364

Tudor B Ionescu and Sebastian Schlund. 2019. A Participatory Programming Model for Democratizing Cobot Technology in Public and Industrial Fablabs. *Procedia CIRP* 81, (2019), 93–98. https://doi.org/10.1016/j.procir.2019.03.017

Sheila Jasanoff and Sang-Hyun Kim. 2009. Containing the Atom: Sociotechnical Imaginaries and Nuclear Power in the United States and South Korea. *Minerva* 47, 2 (June 2009), 119–146. https://doi.org/10.1007/s11024-009-9124-4

Sheila Jasanoff and Sang-Hyun Kim. 2015. *Dreamscapes of Modernity: Sociotechnical Imaginaries and the Fabrication of Power.* University of Chicago Press, Chicago and London. https://doi.org/10.7208/chicago/9780226276663.001.0001

Raya A Jones. 2021. Projective Anthropomorphism as a Dialogue with Ourselves. *International Journal of Social Robotics* (May 2021). https://doi.org/10.1007/s12369-021-00793-7

Matthew Kearnes, Robin Grove-White, Phil Macnaghten, James Wilsdon, and Brian Wynne. 2006. From Bio to Nano: Learning Lessons from the UK Agricultural Biotechnology Controversy. *Science as Culture* 15, 4 (December 2006), 291–307. https://doi.org/10.1080/09505430601022619

Asle H Kiran, Nelly Oudshoorn, and Peter-Paul Verbeek. 2015. Beyond checklists: toward an ethical-constructive technology assessment. *Journal of Responsible Innovation* 2, 1 (January 2015), 5–19. https://doi.org/10.1080/23299460.2014.992769

Asle H Kiran and Peter-Paul Verbeek. 2010. Trusting Our Selves to Technology. *Knowledge, Technology & Policy* 23, 3–4 (December 2010), 409–427. https://doi.org/10.1007/s12130-010-9123-7

Helen Kirkpatrick. 2008. A Narrative Framework for Understanding Experiences of People With Severe Mental Illnesses. *Archives of Psychiatric Nursing* 22, 2 (April 2008), 61–68. https://doi.org/10.1016/j.apnu.2007.12.002

Ronald Kline and Trevor Pinch. 1996. Users as Agents of Technological Change: The Social Construction of the Automobile in the Rural United States. *Technology and Culture* 37, 4 (October 1996), 763. https://doi.org/10.2307/3107097

Bert-Jaap Koops, Angela Di Carlo, Luca Nocco, Vincenzo Casamassima, and Elettra Stradella. 2013. Robotic Technologies and Fundamental Rights: Robotics Challenging the European Constitutional Framework. *International Journal of Technoethics* 4, 2 (July 2013), 15–35. https://doi.org/10.4018/jte.2013070102

John D Lee and Katrina A See. 2004. Trust in Automation: Designing for Appropriate Reliance. *Human Factors* (2004), 31.

Ronald Leenes and Federica Lucivero. 2014. Laws on Robots, Laws by Robots, Laws in Robots: Regulating Robot Behaviour by Design. *Law, Innovation and Technology* 6, 2 (December 2014), 193–220. https://doi.org/10.5235/17579961.6.2.193

Ronald Leenes, Erica Palmerini, Bert-Jaap Koops, Andrea Bertolini, Pericle Salvini, and Federica Lucivero. 2017. Regulatory challenges of robotics: some guidelines for addressing legal and ethical issues. *Law, Innovation and Technology* 9, 1 (January 2017), 1–44. https://doi.org/10.1080/17579961.2017.1304921

Michael Lewis, Katia Sycara, and Phillip Walker. 2018. The Role of Trust in Human-Robot Interaction. In *Foundations of Trusted Autonomy. Studies in Systems, Decision and Control,* Abbass, H., Scholz, J., Reid, D. (eds) 135–159. https://doi.org/10.1007/978-3-319-64816-3_8

Dingjun Li, PL Patrick Rau, and Ye Li. 2010. A cross-cultural study: Effect of robot appearance and task. I*nternational Journal of Social Robotics* 2, 2 (2010), 175–186. https://doi.org/10.1007/s12369-010-0056-9

Patrick Lin, Keith Abney, and George A. Bekey (Eds.). 2012. *Robot ethics: the ethical and social implications of robotics.* MIT Press, Cambridge, Mass.

Hin-Yan Liu and Karolina Zawieska. 2020. From responsible robotics towards a human rights regime oriented to the challenges of robotics and artificial intelligence. Ethics and Information Technology 22, 4 (December 2020), 321–333. https://doi.org/10.1007/s10676-017-9443-3

Lydia He Liu. 2010. *The Freudian robot: digital media and the future of the unconscious.* University of Chicago Press, Chicago.

Arne Maibaum, Andreas Bischof, Jannis Hergesell, and Benjamin Lipp. 2022. A critique of robotics in health care. *AI & Society* 37, 2 (June 2022), 467–477. https://doi.org/10.1007/s00146-021-01206-z

Bertram F Malle. 2016. Integrating robot ethics and machine morality: the study and design of moral competence in robots. *Ethics and Information Technology* 18, 4 (December 2016), 243–256. https://doi.org/10.1007/s10676-015-9367-8

Bertram F Malle, Paul Bello, and Matthias Scheutz. 2019. Requirements for an Artificial Agent with Norm Competence. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (AIES '19), Association for Computing Machinery, New York, NY, USA,* 21–27. https://doi.org/10.1145/3306618.3314252

Mark K. McBeth and Donna L. Lybecker. 2018. The Narrative Policy Framework, Agendas, and Sanctuary Cities: The Construction of a Public Problem. *Policy Studies Journal* 46, 4 (2018), 868–893. DOI:https://doi.org/10.1111/psj.12274

Paul K. McClure. 2018. "You're Fired," Says the Robot: The Rise of Automation in the Workplace, Technophobes, and Fears of Unemployment. *Social Science Computer Review* 36, 2 (April 2018), 139–156. https://doi.org/10.1177/0894439317698637

Michael Nagenborg, Rafael Capurro, Jutta Weber, and Christoph Pingel. 2008. Ethical regulations on robotics in Europe. *AI & Society* 22, 3 (January 2008), 349–366. https://doi.org/10.1007/s00146-007-0153-y

Stanislava Naneva, Marina Sarda Gou, Thomas L. Webb, and Tony J. Prescott. 2020. A Systematic Review of Attitudes, Anxiety, Acceptance, and Trust Towards Social Robots. *International Journal of Social Robotics* 12, 6 (December 2020), 1179–1201. https://doi.org/10.1007/s12369-020-00659-4

Cristopher Nash. 1994. *Narrative in culture: the uses of storytelling in the sciences, philosophy, and literature*. Routledge, London; New York.

Jake Nelson and Tim Gorichanaz. 2019. Trust as an ethical value in emerging technology governance: The case of drone regulation. *Technology in Society* 59, (November 2019), 101131. https://doi.org/10.1016/j.techsoc.2019.04.007

Tatsuya Nomura, Suzuki Tomohiro and Kanda Takayuki. 2006. Altered Attitudes of People toward Robots: Investigation through the Negative Attitudes toward Robots Scale. In *AAAI 2006 workshop on human implications of human-robot interaction,* 7. The AAAI Press, Menlo Park, CA.

Illah Reza Nourbakhsh. 2013. *Robot futures.* The MIT Press, Cambridge, Massachusetts.

Sven Nyholm. 2018. Attributing Agency to Automated Systems: Reflections on Human–Robot Collaborations and Responsibility-Loci. *Science and Engineering Ethics* 24, 4 (August 2018), 1201–1219. https://doi.org/10.1007/s11948-017-9943-x

OECD. 2019. *Recommendation of the Council on Artificial Intelligence*.

Ugo Pagallo. 2010. Robotrust and Legal Responsibility. *Knowledge, Technology & Policy* 23, 3–4 (December 2010), 367–379. https://doi.org/10.1007/s12130-010-9120-x

Nicolas Petit. 2017. *Law and Regulation of Artificial Intelligence and Robots: Conceptual Framework and Normative Implications*. Retrieved October 7, 2019 from https://papers.ssrn.com/abstract=2931339

Michelle M E van Pinxteren, Ruud W H Wetzels, Jessica Rüger, Mark Pluymaekers, and Martin Wetzels. 2019. Trust in humanoid robots: implications for services marketing.

Journal of Services Marketing 33, 4 (January 2019), 507–518. https://doi.org/10.1108/JSM-01-2018-0045

Arjanna van der Plas, Martijntje Smits, and Caroline Wehrmann. 2010. Beyond Speculative Robot Ethics: A Vision Assessment Study on the Future of the Robotic Caretaker. *Accountability in Research* 17, 6 (November 2010), 299–315. https://doi.org/10.1080/08989621.2010.524078

Ibo van de Poel. 2008. How Should We Do Nanoethics? A Network Approach for Discerning Ethical Issues in Nanotechnology. *Nanoethics* 2, 1 (April 2008), 25–38. https://doi.org/10.1007/s11569-008-0026-y

Ibo van de Poel and Lambèr Royakkers. 2011. *Ethics, Technology, and Engineering.* Wiley-Blackwell, Chichester.

Jennifer Rhee. 2018. *The robotic imaginary: The human and the price of dehumanized labor.* University of Minnesota Press, Minneapolis.

Kathleen Richardson. 2015. *An Anthropology of Robots and AI: Annihilation Anxiety and Machines. Routledge.* Routledge. https://doi.org/10.4324/9781315736426

Kathleen Richardson. 2019. The Business of Ethics, Robotics, and Artificial Intelligence. In *Cyborg Futures*, Teresa Heffernan (ed.). Springer, Berlin, 113–126.

Laurel D Riek. 2012. Wizard of Oz studies in HRI: a systematic review and new reporting guidelines. *Journal of Human-Robot Interaction* 1, 1 (July 2012), 119–136. https://doi.org/10.5898/JHRI.1.1.Riek

Kjetil Rommetveit, Niels van Dijk, and Kristrún Gunnarsdóttir. 2020. Make Way for the Robots! Human- and Machine-Centricity in Constituting a European Public–Private Partnership. *Minerva* 58, 1 (March 2020), 47–69. https://doi.org/10.1007/s11024-019-09386-1

Kjetil Rommetveit and Brian Wynne. 2017. Technoscience, imagined publics and public imaginations. *Public Understanding of Science* 26, 2 (February 2017), 133–147. https://doi.org/10.1177/0963662516663057

Francesca Rossi. 2018. Building trust in artificial intelligence. *Journal of international affairs* 72, 1 (2018), 127–134.

Mark Ryan. 2020. In AI We Trust: Ethics, Artificial Intelligence, and Reliability. *Science and Engineering Ethics* 26, 5 (October 2020), 2749–2767. https://doi.org/10.1007/s11948-020-00228-y

Selma Sabanovic. 2010. Robots in Society, Society in Robots: Mutual Shaping of Society and Technology as a Framework for Social Robot Design. *International Journal of Social Robotics* 2, 4 (December 2010), 439–450. https://doi.org/10.1007/s12369-010-0066-7

Henrik Skaug Sætra, Mark Coeckelbergh, and John Danaher. 2021. The AI ethicist's dilemma: fighting Big Tech by supporting Big Tech. *AI Ethics* (December 2021). https://doi.org/10.1007/s43681-021-00123-7

Ziauddin Sardar. 2010. Welcome to postnormal times. *Futures* 42, 5 (June 2010), 435–444. https://doi.org/10.1016/j.futures.2009.11.028

Petra Schaper-Rinkel. 2013. The role of future-oriented technology analysis in the governance of emerging technologies: The example of nanotechnology. *Technological Forecasting and Social Change* 80, 3 (March 2013), 444–452. https://doi.org/10.1016/j.techfore.2012.10.007

Cynthia Selin. 2008. The Sociology of the Future: Tracing Stories of Technology and Time. *Sociology Compass* 2, 6 (November 2008), 1878–1895. https://doi.org/10.1111/j.1751-9020.2008.00147.x

Scott Selisker. 2016. *Human programming: brainwashing, automatons, and American unfreedom.* University of Minnesota Press, Minneapolis.

Kaethe Selkirk, Cynthia Selin, and Ulrike Felt. 2018. A Festival of Futures: Recognizing and Reckoning Temporal Complexity in Foresight. In *Handbook of Anticipation,* Roberto Poli (ed.). Springer International Publishing, Cham, 1–23. https://doi.org/10.1007/978-3-319-31737-3_107-2

Keng Siau and Weiyu Wang. 2018. Building Trust in Artificial Intelligence, Machine Learning, and Robotics. *Cutter Business Technology Journal* 31, 2 (2018), 8.

Jeffry A. Simpson. 2007. Foundations of interpersonal trust. In *Social psychology: Handbook of basic principles,* A. W. Kruglanski & E. T. Higgins (Eds.). The Guilford Press (2007).

Sergio Sismondo. 2020. Sociotechnical imaginaries: An accidental themed issue. *Social Studies of Science* 50, 4 (August 2020), 505–507. https://doi.org/10.1177/0306312720944753

Fran Soljacic, Meia Chita-Tegmark, Theresa Law, and Matthias Scheutz. 2022. *Robots in healthcare as envisioned by care professionals.* https://doi.org/10.48550/arXiv.2206.00776

Bernd Carsten Stahl and Mark Coeckelbergh. 2016. Ethics of healthcare robotics: Towards responsible research and innovation. *Robotics and Autonomous Systems* 86, (December 2016), 152–161. https://doi.org/10.1016/j.robot.2016.08.018

Lucy Suchman. 2006. *Human–Machine Reconfigurations: Plans and Situated Actions.* Cambridge University Press.

Lucy Suchman. 2019. Demystifying the Intelligent Machine. In *Cyborg futures: cross-disciplinary perspectives on artificial intelligence and robotics*, Theresa Heffernan (ed.). Springer, Berlin, 35–61.

Jim Torresen. 2018. A Review of Future and Ethical Perspectives of Robotics and AI. *Frontiers in Robotics and AI* 4, (January 2018), 75. https://doi.org/10.3389/frobt.2017.00075

Jacob Turner. 2019. *Robot rules: regulating artificial intelligence.* Palgrave Macmillan, Basingstoke.

Daniel Ullman and Bertram F Malle. 2018. What Does it Mean to Trust a Robot?: Steps Toward a Multidimensional Measure of Trust. In *Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction - HRI '18,* ACM Press, Chicago, IL, USA, 263–264. https://doi.org/10.1145/3173386.3176991

Steven Umbrello and Angelo Frank De Bellis. 2018. A Value-Sensitive Design Approach to Intelligent Agents. In A*rtificial Intelligence Safety and Security* (2018), Roman Yampolskiy (.ed). CRC Press. Retrieved October 23, 2021 from https://papers.ssrn.com/abstract=3105597

UNESCO. 2021. *DRAFT Recommendation on the Ethics of Artificial Intelligence.* Retrieved from https://unesdoc.unesco.org/ark:/48223/pf0000379920

Núria Vallès-Peris and Miquel Doménech. 2020. Roboticists' Imaginaries of Robots for Care: The Radical Imaginary as a Tool for an Ethical Discussion. *Engineering Studies*, 12(3), (2020), 157–176 https://doi.org/10.1080/19378629.2020.1821695

Adelheid Voskuhl. 2013. *Androids in the Enlightenment: Mechanics, Artisans, and Cultures of the Self.* University of Chicago Press. https://doi.org/10.7208/chicago/9780226034331.001.0001

Astrid Weiss and Katta Spiel. 2021. Robots beyond Science Fiction: mutual learning in human–robot interaction on the way to participatory approaches. *AI & Society* (April 2021). https://doi.org/10.1007/s00146-021-01209-w

Gisa Weszkalnys and Andrew Barry. 2013. Multiple Environments: accountability, integration and ontology. In *Interdisciplinarity: Reconfigurations of the social and natural sciences,* Andrew Barry, Georgina Born (Eds.). Routledge, 194–224.

Martin Willis. 2006. *Mesmerists, Monsters, and Machines: Science Fiction and the Cultures of Science in the Nineteenth Century.* Kent State University Press.

Alan F T Winfield and Marina Jirotka. 2018. Ethical governance is essential to building trust in robotics and artificial intelligence systems. *Philosophical Transactions of the Royal Society A: Mathematical Physical and Engineering Scieinces* 376, 2133 (November 2018), 1–13. https://doi.org/10.1098/rsta.2018.0085

Alan F T Winfield, Katie Winkle, Helena Webb, Ulrik Lyngs, Marina Jirotka, and Carl Macrae. 2021. Robot Accident Investigation: A Case Study in Responsible Robotics. In *Software Engineering for Robotics*, Ana Cavalcanti, Brijesh Dongol, Rob Hierons, Jon Timmis and Jim Woodcock (eds.). Springer International Publishing, Cham, 165–187. https://doi.org/10.1007/978-3-030-66494-7_6

Robert H Wortham and Andreas Theodorou. 2017. Robot transparency, trust and utility. *Connection Science* 29, 3 (July 2017), 242–248. https://doi.org/10.1080/09540091.2017.1313816

Robert H Wortham, Andreas Theodorou, and Joanna J Bryson. 2016. What does the robot think? Transparency as a fundamental design requirement for intelligent systems. In *Proceedings of the IJCAI Workshop on Ethics for Artificial Intelligence: International Joint Conference on Artificial Intelligence.* IJCAI 2016 Ethics for AI Workshop.

Robert H. Wortham, Andreas Theodorou, and Joanna J Bryson. 2017. Robot Transparency: Improving Understanding of Intelligent Behaviour for Designers and Users. In *Towards Autonomous Robotic Systems TAROS 2017. (Lecture Notes in Computer Science)*, Gao, Y., Fallah, S., Jin, Y., Lekakou, C. (eds). Springer International Publishing, Cham, 274–289. https://doi.org/10.1007/978-3-319-64107-2_22

Aimee van Wynsberghe. 2013. Designing Robots for Care: Care Centered Value-Sensitive Design. *Science and Engineering Ethics* 19, 2 (June 2013), 407–433. https://doi.org/10.1007/s11948-011-9343-6

Guang-Zhong Yang, Jim Bellingham, Pierre E Dupont, Peer Fischer, Luciano Floridi, Robert Full, Neil Jacobstein, Vijay Kumar, Marcia McNutt, Robert Merrifield, Bradley J Nelson, Brian Scassellati, Mariarosaria Taddeo, Russell Taylor, Manuela Veloso, Zhong Lin Wang, and Robert Wood. 2018. The grand challenges of Science Robotics. Science Robotics 3, 14 (January 2018), eaar7650. https://doi.org/10.1126/scirobotics.aar7650

Andrej Zwitter. 2014. Big Data ethics. *Big Data & Society* 1, 2 (July 2014), https://doi.org/10.1177/2053951714559253

Robots are increasingly becoming prevalent in our daily lives within our living or working spaces. We hope that robots will take up tedious, mundane or dirty chores and make our lives more comfortable, easy and enjoyable by providing companionship and care. However, robots may pose a threat to human privacy, safety and autonomy; therefore, it is necessary to have constant control over the developing technology to ensure the benevolent intentions and safety of autonomous systems. Building trust in (autonomous) robotic systems is thus necessary.

The title of this book highlights this challenge: “Trust in robots—Trusting robots”. Herein, various notions and research areas associated with robots are unified. The theme “Trust in robots” addresses the development of technology that is trustworthy for users; “Trusting robots” focuses on building a trusting relationship with robots, furthering previous research. These themes and topics are at the core of the PhD program “Trust Robots” at TU Wien, Austria.

ISBN 978-3-85448-051-8

9 783854 480518

www.tuwien.at/academicpress