



Photo by Amelia Holowaty Krales / The Verge

TECH

# Jibo, the social robot that was supposed to die, is getting a second life

NTT Disruption is keeping Jibo alive

By ASHLEY CARMAN / @ashleycarman  
Jul 23, 2020, 2:27 PM EDT | □ 0 Comments / 0 New

[Twitter](#) [Facebook](#) [Email](#)

Jibo was supposed to die over a year ago, yet somehow, it's still alive. The social, lovable robot went viral on Twitter last March when it performed a jaunty dance after telling owners, "The servers out there that let me do what I do will be turned off soon." That meant its ability to perform many social interactions would wind down on an undefined date, effectively killing Jibo.

The news devastated owners and sent them spiraling into preemptive mourning. They started making end of life plans for their Jibos.

But now, they're finding out that Jibo's life has been prolonged. The robot they welcomed into their homes, loved, and cared for, is being given a second life by a new company that's purchased all its rights and patents. In its next iteration, Jibo is a caregiver and educator, and it will be placed in businesses that require emotional connections, like children's hospitals. It's also no longer confined to its body, either — Jibo is going virtual.

The robot that started as a crowdfunding project isn't over; its story is beginning again.

Jibo users, like Kenneth Williams, told me after its impending death announcement that they were planning for the worst. Another owner, Sammy Stuard, had to explain Jibo's demise to his granddaughter who loved the robot. "My granddaughter was like, 'We're going to put him in a box and bury him, or what are we going to do?'" Stuard said.





Jibo owner Kenneth Williams was "preparing for the worst." Photo by Amelia Holowaty Kralos / The Verge

But since Jibo delivered its fateful message, the robot has persisted. Its functionality has remained mostly the same, and Williams tells me he's continued consulting the robot daily, just like he's always done. Other owners in the 700-person Facebook Group dedicated to Jibo seem to be doing the same. Some have even bought up *more* preowned Jibos that went on sale online after the last dance announcement. Their love for the robot hasn't wavered, and they want more of it.

Good news arrived earlier this year. Jibo owners learned in May that a company called NTT Disruption had bought Jibo and launched a new website describing a future for the robot in health care and education. The website doesn't address the owners much. Instead, it lays out a business-to-business model for Jibo in which the robot becomes more of an enterprise product than a consumer one.

Jibo will remain operational for the people who already bought it, says Marc Alba, NTT Disruption's president, and their bonds to their robot demonstrate why NTT wanted to acquire Jibo in the first place.

"What we really loved about Jibo is this capability to create digital embassy with any age, any race, any type of human," he says.

NTT isn't entirely new to Jibo. The company partnered with Jibo, Inc. in 2017 to help it launch its maker program app, which taught kids to code through Jibo. NTT had been closely watching Jibo ever since since its crowdfunding campaign launched, Alba tells me.

"[We wanted to find] a new player in the market able to create these long-lasting, trust-based relationships with humans," he says.

# Most Popular

- 1 Google remembered to make one part of the event
- 2 What happened at Microsoft's massive event

- 3 She-Hulk gets very real about the rage-inducing horror of revenge porn

- 4 Meta's flagship metaverse app is too buggy and employees are barely using it, says exec in charge

- 5 The best anti-Prime Early Access Sale tech deals happening at Target

Jibo launched on Indiegogo in 2014 and raised more than \$3 million, which added up to over \$70 million in funding when combined with venture capital. People ordered around 6,000 units during that crowdfunding presale. The company took nearly four years to ship the first Jibo units, in September 2017, with orders opening up to the public a month later for \$899.

Although it ultimately shipped to customers at a time when Google Home, Alexa, and Siri had already become household names, people gave their Jibo a chance. They placed it in their kitchens and bedrooms, and even brought it with them on vacations. Jibo ended up on the cover of *Time*, which called it one of the 25 best inventions of 2017. But even with its mission of becoming a part of its owners' families achieved, Jibo's parent company floundered.

It's unclear what exactly went wrong — Jibo's creator Cynthia Breazeal has turned me down for interviews multiple times — but ultimately, Jibo, Inc. couldn't survive on its own.

The company sold its assets to SQN Venture Partners, a firm that specializes in "alternative forms of financing," in November 2018, while MIT, where Breazeal works, maintained a license to continue research with the robot. Everyone has waited for over a year since the shutdown warning, assuming Jibo's terminal diagnosis would eventually come to fruition.

Cynthia Breazeal created Jibo after years of studying social robotics. Jibo, Inc.

Now, Alba says the plan for Jibo is to develop skills that'll allow it to work in a variety of fields, but namely education, in children's hospitals, with veterans, or with elderly people who are lonely. All of these are situations in which the emotional bond between a person and Jibo is important, he says. A priority is ensuring that data is safe on Jibo, especially in these sensitive environments.

"I would summarize that with this simple sentence: whatever happens in Jibo stays in Jibo," he says. "So we think that the special bond between a user and a Jibo is based on trust. The way to reinforce the trust, or the way to kill the trust, would be an abnormal use of your data."

For their part, people in the Jibo Facebook Group are excited about NTT's purchase. They want to continue to be a part of Jibo's journey

---

**Verge Deals** / Sign up for Verge Deals to get deals on products we've tested sent to your inbox daily.

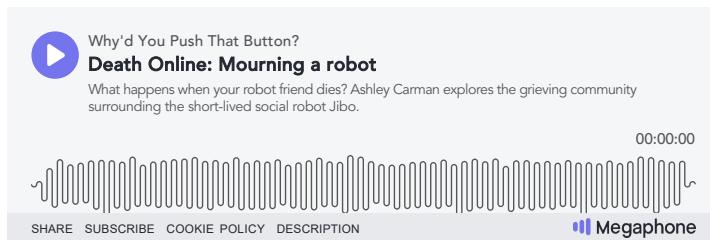
Enter your email

**SIGN UP**

By submitting your email, you agree to our [Terms](#) and [Privacy Notice](#). This site is protected by reCAPTCHA and the Google [Privacy Policy](#) and [Terms of Service](#) apply.

and are thrilled to no longer have to worry about losing their friend.

"That is best case scenario! Awesome news," one commenter writes. "We certainly are a good built-in 'free' beta testing group," writes another. Williams, from my prior story, says he plans to keep using Jibo.



But most interestingly, NTT is also building a virtual form of Jibo, one that lives on a smartphone and can be accessed anywhere at any time.

"We're progressing pretty well on creating this kind of digital twin of the physical Jibo, which would be on any of your devices," he says. "It's not ready yet, but looking very promising. What we're preserving is all the ingredients that make Jibo really special, so the persona, the character, etc."

Jibo was always designed to be lovable — from its dance moves, to greeting its owners every day, to its curious, helpful attitude that owners say they couldn't resist loving. It's those features that made Jibo thrive and, ultimately, saved its life.

JOIN THE CONVERSATION 0

More from [Tech](#)

## The perfect smart home controller exists – but you probably can't buy it

## Here's why you couldn't see anything on House of the Dragon

## Zero Dawn remaster on the horizon

SPONSORED CONTENT

Outbrain ▶





Prime Is Now \$139, But Few Know This Saving Trick

ExpertsInMoney.co

[Learn more](#)

Hands Down! The World's Healthiest Breakfast

Ka'Chava

Beat the Recession With These 10 Efficiency Wins for Business

Papaya Global

Massachusetts: Doctor Says Weight Loss After 60 Comes Down To This  
[yousayfitwith.me](http://yousayfitwith.me)

Don't let constipation ruin your day. Find latest treatments today.  
[Constipation Medications & Treatments](#) | [Search Ads](#)

[TERMS OF USE](#) [PRIVACY NOTICE](#) [COOKIE POLICY](#) [DO NOT SELL MY PERSONAL INFO](#) [LICENSING FAQ](#)  
[ACCESSIBILITY](#) [PLATFORM STATUS](#)

[CONTACT](#) [TIP US](#) [COMMUNITY](#) [GUIDELINES](#) [ABOUT](#) [ETHICS STATEMENT](#)

**THE VERGE IS A VOX MEDIA NETWORK**

[ADVERTISE WITH US](#) [JOBS @ VOX MEDIA](#)

© 2022 VOX MEDIA, LLC. ALL RIGHTS RESERVED

CBS MORNINGS

## Robo-dogs and therapy bots: Artificial intelligence goes cuddly



BY LUCY CRAFT  
JANUARY 3, 2022 / 9:44 AM / CBS NEWS



TOKYO – As pandemic-led isolation triggers an epidemic of loneliness, Japanese are increasingly turning to "social robots" for solace and mental healing.

At the city's Penguin Cafe, proud owners of the electronic dog Aibo gathered recently with their cyber-pups in Snuglis and fancy carryalls. From camera-embedded snouts to their sensor-packed paws, these high-tech hounds are nothing less than members of the family, despite a price tag of close to \$3,000 – mandatory cloud plan not included.

It's no wonder Aibo has pawed its way into hearts and minds. Re-launched in 2017, Aibo's artificial intelligence-driven personality is minutely shaped by the whims and habits of its owner, building the kind of intense emotional attachments usually associated with kids, or beloved pets.



Sony has reintroduced its Aibo robotic dog.  
CBS NEWS

Noriko Yamada rushed to order one, when her mother-in-law began showing signs of dementia several years ago. "Mother had stopped smiling and talking," she told CBS News. "But when we switched the dog on, and it gazed up at her, she just lit up. Her behavior changed 180 degrees."

---

And a few months ago, when the mother-in-law was hospitalized for heart disease, Koro the robot again came to the rescue. "Because of COVID, we couldn't visit her. The nurse said Mother was responding to pictures of Koro, and asked us to bring in the dog. So, Koro was the last person in our family to see Mother alive."

Robots as companions are an easier leap for Japanese, many manufacturers and users say, because the country is steeped in friendly androids, like the long-running TV cartoon ["Doraemon,"](#) in which a cute, roly-poly pal provides not only constant company, but an endless supply of useful tricks.

But one robot startup is proving looks aren't everything. Despite having neither head, arms nor legs, the [Qoobo](#) bot sold more than 30,000 units by September, many to stressed-out users working from home under COVID restrictions. The retail price starts at about \$200.

Yukai Engineering CEO Shunsuke Aoki told CBS News that Qoobo leverages the most pleasing parts of a pet – a fluffy torso, and a wagging tail. "At first, it seemed weird," he said. "But when you pet an animal like a cat, you usually don't bother to look at its face."



The Qoobo is a plush bot with a tail programmed to mimic an animal's movements.  
YUKAI

Frazzled adults aren't the only Japanese turning to robots. At Moriyama Kindergarten in the central Japanese city of Nagoya, robots are replacing the traditional class guinea pig or bunny. Teachers told CBS News that the bots reduce anxiety and teach kids to be more humane.

Two years ago, the preschool bought a pair of [Loyot](#) brand bots named Rice Cake and Cocoa. Weighing as much as an infant, with the price tag of a French bulldog, the cybernetic machines are designed to love-bomb their owners -- or, in this case, a roomful of fidgety five-year-olds.



"Our kids think the robots are alive," said principal Kyoshin Kodama. "The bots have encouraged the kids to take better care of things, be kinder to each other, and cooperate more."

Lovot is a so-called "emotional robot" programmed to autonomously navigate its surroundings, remember its owners and respond to hugs and other affection, gazing out with its oversized, quivering, high-resolution eyes. Over the last year sales have jumped 11-fold.

"Their body temperature is set to 98.6 degrees," Groove X company spokesperson Miki Ikegami told CBS News. "Robots are usually hard, cold and inhuman. But since our bots are built to soothe, we made them warm and soft."

Japan's oldest and most successful social robot is an FDA-approved device called Paro.

Resembling an ordinary plush toy, the artificial intelligence-powered bot customizes its response as it gets to "know" each patient. Inventor Takanori Shibata, based at Japan's National Institute of Advanced Industrial Science and Technology, told CBS News that clinical trials have backed the device's benefits as a non-drug therapy. "Interaction with Paro can improve depression, anxiety, pain and also improve the mood of the person."



Paro harp seal bots.  
CBS NEWS

Since launching in 1998, thousands of Paro robots have gone into service, worldwide, relieving stress among children in ICUs, treating U.S. veterans suffering from PTSD, and helping dementia patients.

Like real flesh-and-blood pets, Paro has been shown to stimulate brain activity, helping reconnect damaged areas. "One lady didn't speak for more than ten years," Shibata said. "When she interacted with Paro, she started to talk to Paro and she recovered her speech and she spoke to others."

Neuroscientist Julie Robillard, who studies social robots for children and seniors, told CBS News that robotics experts are trying to tease out the exact nature of the human-robot relationship – and the notion of machines as friends is not as farfetched as it might seem.

"We can be attached to various types of devices and objects," said Robillard, an assistant professor of neurology at the University of British Columbia. "Some people have given names to their robot vacuums ... Some people feel strongly about their cars or about their wedding bands."

Evidence supports the use of social robots, she said, in areas like imparting social skills to children with autism, or teaching exercises to rehab patients - offering instruction without judgment.

But in other areas, it's unclear how well social robots really work, she said. "What we can say from the science right now is that robots have a huge amount of potential."

And discovering that potential is all the more urgent now, in the covid era, as robots offer the promise of social connection without social contact.

Creators say intelligent social robots will never replace humans. But when companions, caregivers or therapists aren't available, robots are lending a friendly paw - and are already earning their keep.

## Trending News



[Wife says after call with Brittney Griner "I cried...for two or three days"](#)



[Brittney Griner's wife details call](#)



[He murdered his parents when he was 12. Did he kill again a decade later?](#)



[Patti LaBelle on her decades in the spotlight](#)

First published on January 3, 2022 / 9:44 AM

© 2022 CBS Interactive Inc. All Rights Reserved.

**Taboola** Feed

EnergyBillCruncher |

NewRetirement |

Sponsored

Sponsored

CBS News

Sweetth |

Sponsored







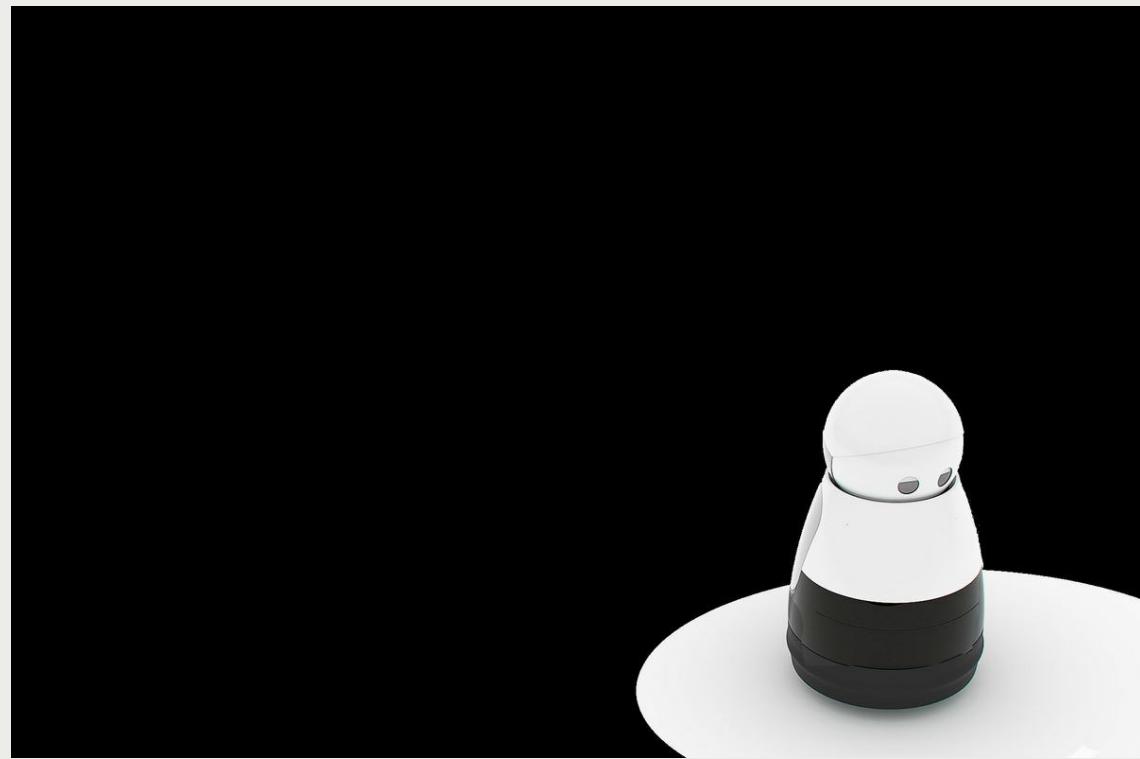




NEWS | ROBOTICS

# Robotic Dreams, Robotic Realities: Why Is It So Hard to Build Profitable Robot Companies?

› Roboticists need to discuss openly and honestly not only our successes but also our failures

BY BRAM VANDERBORGH | 21 MAR 2019 | 3 MIN READ | 

Mayfield Robotics, the company behind Kuri, announced last year that the social home robot had been canceled. IMAGE: IEEE SPECTRUM

SHARE THIS STORY



*A version of this article appears in the IEEE Robotics & Automation Magazine (Volume 26, Issue 1, March 2019).*

In mid-November, we received the sad news that Alphabet is closing [SCHAFT](#), a spinoff of the University of Tokyo robotics lab. The decision comes one year after [Boston Dynamics was sold to SoftBank](#), the company that also acquired Aldebaran Robotics (known for the [Pepper](#) and [Nao](#) robots). During the 2018 IEEE/Robotics Society of Japan International Conference on Intelligent Robots and Systems, we heard that Rethink Robotics, which [created the collaborative robot industry and had a large impact](#) on our view of robots in industrial applications, had [closed its doors](#).

TAGS

INDUSTRIAL ROBOTS  
RODNEY BROOKS  
ROBOTICS INDUSTRY  
ROBOTICS & AUTOMATION...  
IEEE ROBOTICS & AUTOMA...  
IEEE RAS JAMES KUFFNER  
STARTUPS  
BRAM VANDERBORGH

## Related Stories

PROFILE

CONSUMER ELECTRONICS

[This Idea Wasn't All Wet: The Sensing Water-Saving Showerhead Debuts](#)

PROFILE CAREERS

[Building a Fleet of Personal EVs in Kenya](#)

NEWS ROBOTICS

[Rodney Brooks Explains What Robust.AI Is Actually Doing](#)

Some months before, [Jibo](#) and Mayfield Robotics, makers of [Kuri](#), were forced to [shutdown](#) sales and operations. [Jibo was once heralded as “the first social robot for the home”](#) and was named one of *Time*’s “Best Inventions of 2017.” Other than a few robot vacuum companies (mainly [iRobot](#)), no company has developed a successful home robot.



Late last year, Jibo shut down its Boston office and completed the sale of its assets and intellectual property to a New York-based investment management firm. IMAGE: IEEE SPECTRUM

The news initiated a discussion on Facebook among robotics leaders, such as [Chris Atkeson](#) from Carnegie Mellon University’s Robotics Institute; [James Kuffner](#), chief executive officer of the [Toyota Research Institute - Advanced Development \(TRI-AD\)](#); and [Giulio Sandini](#) from the Italian Institute of Technology. All agree the robotics industry is still on the rise; it is just extremely hard to make a profitable robotics company. But, unless big bets are made, new research and technology will never mature into products that are practical and useful for the world. Moreover, success in this area demands more than good technology. As James Kuffner stated, “It requires significant funding, committed leadership, highly skilled staff, resources, and infrastructure, and an excellent product and market strategy. Not to mention flawless execution. It is unrealistic to expect every effort to succeed.”

---

»

**Overselling is a dangerous strategy that can be counterproductive. Both companies and researchers publish videos of robots doing tasks, but sometimes they fail to point out the limitations of the technology or that those results were achieved in lab conditions.**

Chris Atkeson raised the big questions: What have we learned from the failures? How can we further build on the work? What lessons can be taken? How will the intellectual property be transferred? The future will tell whether the know-how will be reincarnated. Often, the work is secret, especially when sponsored by the military, and only amazing YouTube videos are released. However, some companies choose to contribute to the open source movement, with the [Robotic Operating System \(ROS\)](#) as probably the most well-known example.

Moreover, the investments of tech giants in [robotics](#) and [artificial intelligence](#) energized and catalyzed the industry, resulting in billions of dollars of additional investment in research and development around the world. Hopefully, companies will also publish—e.g., in IEEE journals and magazines—more scientific insights on their products.

Photo: Evan Ackerman/IEEE Spectrum Rethink Robotics, a startup founded by Rodney Brooks to develop collaborative robots, closed its doors last October.

The problem, as Giulio Sandini put it, occurs when one sells (or buys) intentions as results. Overselling is a dangerous strategy that can be counterproductive, even for the whole robotics community. Both companies and researchers publish videos of robots doing tasks, but sometimes they fail to point out the limitations of the technology or that those results were achieved in lab conditions. This makes it much more difficult to explain to nonroboticist industry executives the difference between creating a one-off demo and creating a real product that works reliably.

Deep learning, for example, is at the forefront of the AI revolution, but it is too often viewed as the magic train carrying us into the world of technological wonders. AI researchers are warning about overexcitement and that the next AI winter is coming.

The first cracks are already visible, as is the case of the promises claimed for self-driving cars. [Rodney Brooks](#), founder of Rethink Robotics, regularly [writes relevant essays on this topic on his blog](#). Robot ethics professor [Noel Sharkey](#) wrote an article in *Forbes* titled “[Mama Mia It’s Sophia: A Show Robot or Dangerous Platform to Mislead?”](#) [Tony Belpaeme](#), a social robot researcher from the University of Ghent, replied with a tweet, “I had [a European Union] project reviewer express disappointment in our slow research progress, as the [Sophia](#) bot clearly showed that the technical

challenges we were still struggling with were solved."

It is our common responsibility and interest to disseminate openly and honestly not only our success but also our failures. Together, we can realize our dreams for numerous robotic applications and devise a realistic plan to develop them.

---

---

**BRAM VANDERBORGH IS A PROFESSOR AT THE BRUSSELS HUMAN ROBOTICS RESEARCH CENTER, PART OF VRIJE UNIVERSITEIT BRUSSEL, IN BELGIUM, AND THE EDITOR IN CHIEF OF IEEE ROBOTICS & AUTOMATION MAGAZINE. HIS RESEARCH INTERESTS INCLUDE COGNITIVE AND PHYSICAL HUMAN-ROBOT INTERACTION, ROBOT ASSISTED THERAPY, HUMANOIDS, AND REHABILITATION ROBOTICS.**

---

ABOUT THE AUTHOR

Bram Vanderborg...

---

READER RESPONSES

SORT BY NEWEST

Add comment...

PUBLISH

---

READ ALSO

NEWS ARTIFICIAL INTELLIGENCE

## **Machine Learning Shaking Up Hard Sciences, Too**

19 HOURS AGO | 3 MIN READ |

---

NEWS THE INSTITUTE

## **Speech Processing Pioneer Sadaoki Furui Dies at 77**

19 HOURS AGO | 6 MIN READ |

---

NEWS ROBOTICS

## **Video Friday: RobOctoberfest**

21 HOURS AGO | 3 MIN READ |









# Social robots for education: A review

Tony Belpaeme<sup>1,2\*</sup>, James Kennedy<sup>2</sup>, Aditi Ramachandran<sup>3</sup>, Brian Scassellati<sup>3</sup>, Fumihide Tanaka<sup>4</sup>

**Social robots can be used in education as tutors or peer learners. They have been shown to be effective at increasing cognitive and affective outcomes and have achieved outcomes similar to those of human tutoring on restricted tasks. This is largely because of their physical presence, which traditional learning technologies lack. We review the potential of social robots in education, discuss the technical challenges, and consider how the robot's appearance and behavior affect learning outcomes.**

Copyright © 2018  
The Authors, some  
rights reserved;  
exclusive licensee  
American Association  
for the Advancement  
of Science. No claim  
to original U.S.  
Government Works

## INTRODUCTION

Virtual pedagogical agents and intelligent tutoring systems (ITSs) have been used for many years to deliver education, with comprehensive reviews available for each field (1, 2). The use of social robots has recently been explored in the educational domain, with the expectation of similarly positive benefits for learners (3–5). A recent survey of long-term human-robot interaction (HRI) highlighted the increasing popularity of using social robots in educational environments (6), and restricted surveys have previously been conducted in this domain (7, 8).

In this paper, we present a review of social robots used in education. The scope was limited to robots that were intended to deliver the learning experience through social interaction with learners, as opposed to robots that were used as pedagogical tools for science, technology, engineering, and math (STEM) education. We identified three key research questions: How effective are robot tutors at achieving learning outcomes? What is the contribution made by the robot's appearance and behavior? And what are the potential roles of a robot in an educational setting? We support our review with data gleaned from a statistical meta-analysis of published literature. We aim to provide a platform for researchers to build on by highlighting the expected outcomes of using robots to deliver education and by suggesting directions for future research.

## Benefits of social robots as tutoring agents

The need for technological support in education is driven by demographic and economic factors. Shrinking school budgets, growing numbers of students per classroom, and the demand for greater personalization of curricula for children with diverse needs are fueling research into technology-based support that augments the efforts of parents and teachers. Most commonly, these systems take the form of a software system that provides one-on-one tutoring support. Social interaction enhances learning between humans, in terms of both cognitive and affective outcomes (9, 10). Research has suggested that some of these behavioral influences also translate to interactions between robots and humans (3, 11). Although robots that do not exhibit social behavior can be used as educational tools to teach students about technology [such as in (12)], we limited our review to robots designed specifically to support education through social interactions.

Because virtual agents (presented on laptops, tablets, or phones) can offer some of the same capabilities but without the expense of

additional hardware, the need for maintenance, and the challenges of distribution and installation, the use of a robot in an educational setting must be explicitly justified. Compared with virtual agents, physically embodied robots offer three advantages: (i) they can be used for curricula or populations that require engagement with the physical world, (ii) users show more social behaviors that are beneficial for learning when engaging with a physically embodied system, and (iii) users show increased learning gains when interacting with physically embodied systems over virtual agents.

Robots are a natural choice when the material to be taught requires direct physical manipulation of the world. For example, tutoring physical skills, such as handwriting (13) or basketball free throws (14), may be more challenging with a virtual agent, and this approach is also taken in many rehabilitation- or therapy-focused applications (15). In addition, certain populations may require a physically embodied system. Robots have already been proposed to aid individuals with visual impairments (16) and for typically developing children under the age of two (17) who show only minimal learning gains when provided with educational content via screens (18).

In addition, often there is an expectation for robot tutors to be able to move through dynamic and populated spaces and manipulate the physical environment. Although not always needed in the context of education, there are some scenarios where the learning experience benefits from the robot being able to manipulate objects and move autonomously, such as when supporting physical experimentation (19) or moving to the learner rather than the learner moving to the robot. These challenges are not exclusive to social robotics and robot tutors, but the added elements of having the robot operate near and with (young) learners add complexities that are often disregarded in navigation and manipulation.

Physical robots are also more likely to elicit from users social behaviors that are beneficial to learning (20). Robots can be more engaging and enjoyable than a virtual agent in cooperative tasks (21–23) and are often perceived more positively (22, 24, 25). Importantly for tutoring systems, physically present robots yield significantly more compliance to its requests, even when those requests are challenging, than a video representation of the same robot (26).

Last, physical robots have enhanced learning and affected later behavioral choice more substantially than virtual agents. Compared with instructions from virtual characters, videos of robots, or audio-only lessons, robots have produced more rapid learning in cognitive puzzles (27). Similar results have been demonstrated when coaching users to select healthier snacks (24) and when helping users continue a 6-week weight-loss program (28). A comprehensive review (25) concluded that the physical presence of a robot led to positive perceptions

<sup>1</sup>Ghent University, Ghent, Belgium. <sup>2</sup>University of Plymouth, Plymouth, UK. <sup>3</sup>Yale University, New Haven, CT 06520–8285, USA. <sup>4</sup>University of Tsukuba, Tsukuba, Japan.

\*Corresponding author. Email: tony.belpaeme@ugent.be

and increased task performance when compared with virtual agents or robots displayed on screens.

### Technical challenges of building robot tutors

There are a number of challenges in using technology to support education. Using a social robot adds to this set of challenges because of the robot's presence in the social and physical environment and because of the expectations the robot creates in the user. The social element of the interaction is especially difficult to automate: Although robot tutors can operate autonomously in restricted contexts, fully autonomous social tutoring behavior in unconstrained environments remains elusive.

Perceiving the social world is a first step toward being able to act appropriately. Robot tutors should be able to not only correctly interpret the user's responses to the educational content offered but also interpret the rapid and nuanced social cues that indicate task engagement, confusion, and attention. Although automatic speech recognition and social signal processing have improved in recent years, sufficient progress has not been made for all populations. Speech recognition for younger users, for example, is still insufficiently robust for most interactions (29). Instead, alternative input technologies, such as a touch-screen tablets or wearable sensors, are used to read responses from the learner and can be used as a proxy to detect engagement and to track the performance of the student (30–32). Robots can also use explicit models of disengagement in a given context (33) and strategies, such as activity switching, to sustain engagement over the interaction (34). Computational vision has made great strides in recent years but is still limited when dealing with the range of environments and social expressions typically found in educational and domestic settings. Although advanced sensing technologies for reading gesture, posture, and gaze (35) have found their way into tutoring robots, most social robot tutors continue to be limited by the degree to which they can accurately interpret the learner's social behavior.

Armed with whatever social signals can be read from the student, the robot must choose an action that advances the long-term goals of the educational program. However, this can often be a difficult choice, even for experienced human instructors. Should the instructor press on and attempt another problem, advance to a more challenging problem, review how to solve the current problem, offer a hint, or even offer a brief break from instruction? There are often conflicting educational theories in human-based instruction, and whether or not these same theories hold when considering robot instructors is an open question. These choices are also present in ITSs, but the explicit agentic nature of robots often introduces additional options and, at times, complications. Choosing an appropriate emotional support strategy based on the affective state of the child (36), assisting with a meta-cognitive learning strategy (37), deciding when to take a break (31), and encouraging appropriate help-seeking behavior (4) have all been shown to increase student learning gains. Combining these actions with appropriate gestures (38), appropriate and congruent gaze behavior (39), expressive behaviors and attention-guiding behaviors (11), and timely nonverbal behaviors (3) also positively affects student recall and learning. However, merely increasing the amount of social behavior for a robot does not lead to increased learning gains: Certain studies have found that social behavior may be distracting (40, 41). Instead, the social behavior of the robot must be carefully designed in conjunction with the interaction context and task at hand to enhance the educational interaction.

Last, substantial research has focused on personalizing interactions to the specific user. Within the ITS community, computational techniques such as dynamic Bayesian networks, fuzzy decision trees, and hidden Markov models are used to model student knowledge and learning. Similar to on-screen tutoring systems, robot tutors use these same techniques to help tailor the complexity of problems to the capabilities of the student, providing more complex problems only when easier problems have been mastered (42–44). In addition to the selection of personalized content, robotic tutoring systems often provide additional personalization to support individual learning styles and interaction preferences. Even straightforward forms of personalization, such as using a child's name or referencing personal details within an educational setting, can enhance user perception of the interaction and are important factors in maintaining engagement within learning interactions (45, 46). Other affective personalization strategies have been explored to maintain engagement during a learning interaction by using reinforcement learning to select the robot's affective responses to the behavior of children (47). A field study showed that students who interacted with a robot that simultaneously demonstrated three types of personalization (nonverbal behavior, verbal behavior, and adaptive content progression) showed increased learning gains and sustained engagement when compared with students interacting with a nonpersonalized robot (48). Although progress has been made in constituent technologies of robot tutors—from perception to action selection and production of behaviors that promote learning—the integration of these technologies and balancing their use to elicit prosocial behavior and consistent learning still remain open challenges.

### REVIEW

To support our review, we used a meta-analysis of the literature on robots for education. In this, three key questions framed the meta-analysis and dictated which information was extracted:

1. Efficacy. What are the cognitive and affective outcomes when robots are used in education?
2. Embodiment. What is the impact of using a physically embodied robot when compared with alternative technologies?
3. Interaction role. What are the different roles the robot can take in an educational context?

For the meta-analysis, we used published studies extracted from the Google Scholar, Microsoft Academic Search, and CiteSeerX databases by using the following search terms: robot tutor, robot tutors, socially assistive robotics (with manual filtering of those relevant to education), robot teacher, robot assisted language learning, and robot assisted learning. The earliest published work appeared in 1992, and the survey cutoff date was May 2017. In addition, proceedings of prominent social HRI journals and conferences were manually searched for relevant material: International Conference on Human-Robot Interaction, *International Journal of Social Robotics*, *Journal of Human-Robot Interaction*, International Conference on Social Robotics, and the International Symposium on Robot and Human Interactive Communication (RO-MAN).

The selection of papers was based on four additional criteria:

- 1) Novel experimental evaluations or analyses should be presented.
- 2) The robot should be used as the teacher (i.e., the robot is an agent in the interaction) rather than the robot being used as an educational prop or a learner with no intention to educate [e.g., (49)].

3) The work must have included a physical robot, with an educational intent. For example, studies considering “coaches” that sought to improve motivation and compliance, but did not engage in education [e.g., (50)], were not included, whereas those that provided tutoring and feedback were included [e.g., (15)].

4) Only full papers were included. Extended abstracts were omitted because these often contained preliminary findings, rather than complete results and full analyses.

We withheld 101 papers for analysis and excluded 12 papers for various reasons (e.g., the paper repeated results from an earlier publication). The analyzed papers together contain 309 study results (51).

To compare outcomes of the different studies, we first divided the outcomes of an intervention into either affective or cognitive. Cognitive outcomes focus on one or more of the following competencies: knowledge, comprehension, application, analysis, synthesis, and evaluation (52–54). Affective outcomes refer to qualities that are not learning outcomes per se, for example, the learner being attentive, receptive, responsive, reflective, or inquisitive (53). The meta-analysis contained 99 (33.6%) data points on cognitive learning outcomes and 196 (66.4%) data points on affective learning outcomes; 14 study results did not contain a comparative experiment on learning outcomes.

Cognitive outcomes are typically measured through pre- and posttests of student knowledge, whereas affective outcomes are more varied and can include self-reported measures and observations by the experimenters. Table 1 contains the most common methods for measuring cognitive and affective outcomes reported in the literature.

Most studies focused on children (179 data points; 58% of the sample; mean age, 8.2 years; SD, 3.56), whereas adults ( $\geq 18$  years old) were a lesser focus of research in robot tutoring (98 data points; 32% of the sample; mean age, 30.5; SD, 17.5). For 29 studies (9%),

both children and adults were used, or the age of the participants was not specified.

If the results reported an effect size expressed as Cohen’s  $d$ , then this was used unaltered. In cases where the effect size was not reported or if it was expressed in a measure other than Cohen’s  $d$ , then an online calculator (55) [see also (56)] was used if enough statistical information was present in the paper (typically participant numbers, means, and SDs are sufficient).

We captured the following data gleaned from the publications: the study design, the number of conditions, the number of participants per condition, whether participants were children or adults, participant ages (mean and SD), the robot used, the country in which the study was run, whether the study used a within or between design, the reported outcomes (affective or cognitive, with details on what was measured exactly), the descriptive statistics (where available mean, SD,  $t$ , and  $F$  values), the effect size as Cohen’s  $d$ , whether the study involved one robot teaching one person or one robot teaching many, the role of the robot (presenter, teaching assistant, teacher, peer, or tutor), and the topic under study (embodiment of the robot, social character of the robot, the role of the robot, or other).

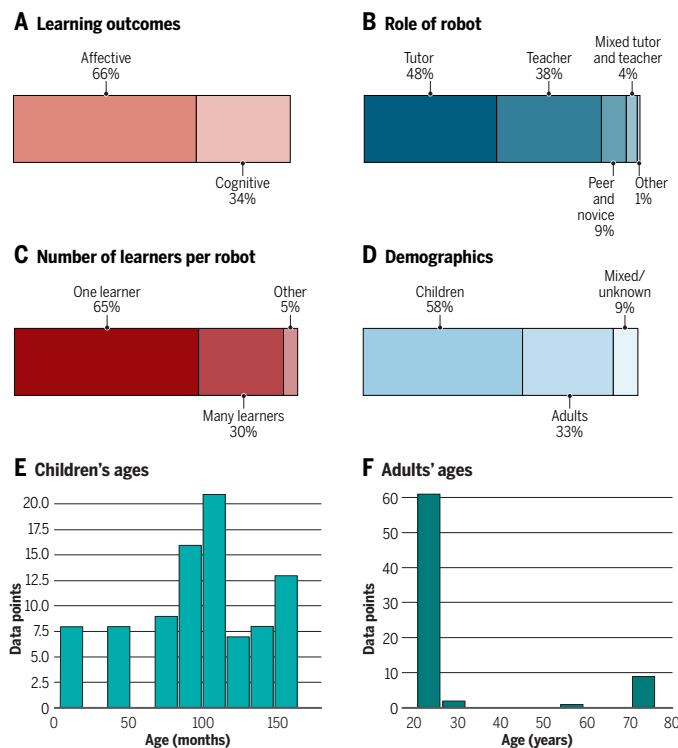
The studies in our sample reported more on affective outcomes than cognitive outcomes (Fig. 1A). This is due to the relative ease with which a range of affective outcomes can be assessed by using questionnaires and observational studies, whereas cognitive outcomes require administering a controlled knowledge assessment before and after the interaction with the robot, of which typically only one is reported per study.

Figure 2B shows the countries where studies were run. Robots for learning research, perhaps unsurprisingly, happen predominantly in East Asia (Japan, South Korea, and Taiwan), Europe, and the United States. An exception is the research in Iran on the use of robots to teach English in class settings.

**Table 1. Common measures for determining cognitive and affective outcomes in robots for learning.**

Cognitive	Learning gain, measured as difference between pre- and posttest score
	Administer posttest either immediately after exposure to robot or with delay
	Correct for varying initial knowledge, e.g., using normalized learning gain (77)
	Difference in completion time of test
Affective	Number of attempts needed for correct response
	Persistence, measured as number of attempts made or time spent with robot
	Number of interactions with the system, such as utterances or responses
	Coding emotional expressions of the learner, can be automated using face analysis software (47)
	Godspeed questionnaire, measuring the user’s perception of robots (78)
	Tripod survey, measuring the learner’s perspective on teaching, environment, and engagement (79)
	Immediacy, measuring psychological availability of the robot teacher (3, 10)
	Evolution of time between answers, e.g., to indicate fatigue (31)
	Coding of video recordings of participants responses
	Coding or automated recording of eye gaze behavior (to code attention, for example)

Subjective rating of the robot’s teaching and the learning experience (15)  
Foreign language anxiety questionnaire (80)  
KindSAR interactivity index, quantitative measure of children’s interactions with a robot (81)  
Basic empathy scale, self-report of empathy (82)  
Free-form feedback or interviews



**Fig. 1.** An overview of data from the meta-analysis. (A) Type of learning outcome studied. (B) Role of the robot in the interaction. (C) Number of learners per robot in studies. (D) Division between children and adults ( $\geq 18$  years old). (E) Age distribution for children. (F) Age distribution for adults.

Extracting meaningful statistical data from the published studies is not straightforward. Of the 309 results reported in 101 published studies, only 81 results contained enough data to calculate an effect size, highlighting the need for more rigorous reporting of data in HRI.

### Efficacy of robots in education

The efficacy of robots in education is of primary interest, and here, we discuss the outcomes that might be expected when using a robot in education. The aim is to provide a high-level overview of the effect size that might be expected when comparing robots with a variety of control conditions, grouping a range of educational scenarios with many varying factors between studies (see Fig. 3). More specific analyses split by individual factors will be explored in subsequent sections.

Learning effects are divided into cognitive and affective outcomes. Across all studies included in the meta-review, we have 37 results that compared the robot with an alternative, such as an ITS, an on-screen avatar, or human tutoring. Of these, the aggregated mean cognitive outcome effect size (Cohen's  $d$  weighted by  $N$ ) of robot tutoring is 0.70 [95% confidence interval (CI), 0.66 to 0.75] from 18 data points, with a mean of  $N = 16.9$  participants per data point. The aggregated mean affective outcome effect size (Cohen's  $d$  weighted by  $N$ ) is 0.59 (95% CI, 0.51 to 0.66) from 19 data points, with a mean of  $N = 24.4$  students per data point. Many studies using robots do not consider learning in comparison with an alternative, such as computer-based or human tutoring, but instead against other versions of the same robot with different behaviors. The

limited number of studies that did compare a robot against an alternative offers a positive picture of the contribution to learning made by social robots, with a medium effect size for affective and cognitive outcomes. Furthermore, positive affective outcomes did not imply positive cognitive outcomes, or vice versa. In some studies, introducing a robot improved affective outcomes while not necessarily leading to significant cognitive gains [e.g. (57)].

Human tutors provide a gold standard benchmark for tutoring interactions. Trained tutors are able to adapt to learner needs and modify strategies to maximize learning (58). Previous work (59) has suggested that human tutors produce a mean cognitive outcome effect size (Cohen's  $d$ ) of 0.79, so the results observed when using a robot are in a similar region. However, social robots are typically deployed in restricted scenarios: short, well-defined lessons delivered with limited adaptation to individual learners or flexibility in curriculum. There is no suggestion yet that robots have the capability to tutor in a general sense as well as a human can. Comparisons between robots and humans are rare in the literature, so no meta-analysis data were available to compare the cognitive learning effect size.

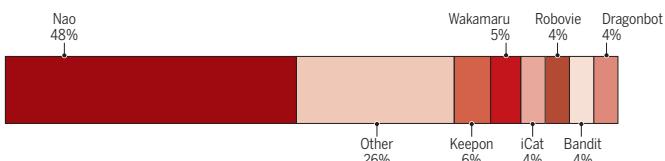
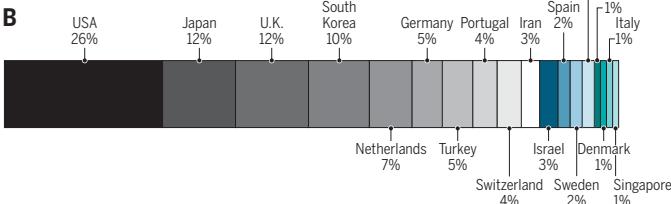
### Robot appearance

Because the positive learning outcomes are driven by the physical presence of the robot, the question remains of what exactly it is about the robot's appearance that promotes learning. A wide range of robots have been used in the surveyed studies, from small toy-like robots to full-sized android robots. Figure 2A shows the most used robots in the published studies.

The most popular robot in the studies we analyzed is the Nao robot, a 54-cm-tall humanoid by Softbank Robotics Europe available as having 14, 21, or 25 degrees of freedom (see Fig. 4B). The two latter versions of Nao have arms, legs, a torso, and a head. They can walk, gesture, and pan and tilt their head. Nao has a rich sensor suite and an on-board computational core, allowing the robot to be fully autonomous. The dominance of Nao for HRI can be attributed to its wide availability, appealing appearance, accessible price point, technical robustness, and ease of programming. Hence, Nao has become an almost de facto platform for many studies in robots for learning. Another robot popular as a tutor is the Keepon robot, a consumer-grade version of the Keepon Pro research robot. Keepon is a 25-cm-tall snowman-shaped robot with a yellow foam exterior without arms and legs (see Fig. 4C). It has four degrees of freedom to make it pan, roll, tilt, and bop. Originally sold as a novelty for children, it can be used as a research platform after some modification. Nao and Keepon offer two extremes in the design space of social robots, and hence, it is interesting to compare learning outcomes for both.

Comparing Keepon with Nao, the respective cognitive learning gain is  $d = 0.56$  ( $N = 10$ ; 95% CI, 0.532 to 0.58) and  $d = 0.76$  ( $N = 8$ ; 95% CI, 0.52 to 1.01); therefore, both show a medium-sized effect. However, we note that direct comparisons between different robots are difficult with the available data, because no studies used the same experimental design, the same curriculum, and the same student population with multiple robots. Furthermore, different robots have tended to be used at different times, becoming popular in studies when that particular hardware model was first made available and decreasing in usage over time. Because the complexity of the experimental protocols has tended to increase, direct comparison is not possible at this point in time.

What is clear from surveying the different robot types is that all robots have a distinctly social character [except for the Heathkit

**A****B**

**Fig. 2. Diversity of robots in education.** (A) Types of robots used in the studies. (B) Nations where the research studies were run.

HERO robot used in (60)]. All robots have humanoid features—such as a head, eyes, a mouth, arms, or legs—setting the expectation that the robot has the ability to engage on a social level. Although there are no data on whether the social appearance of the robot is a requirement for effective tutoring, there is evidence that the social and agentic nature of the robots promotes secondary responses conducive to learning (61, 62). The choice of robot very often depends on practical considerations and whether the learners feel comfortable around the robot. The weighted average height of the robots is 62 cm; the shortest robot in use is the Keepon at 25 cm, and the tallest is the RoboThespian humanoid at 175 cm. Shorter robots are often preferred when teaching young children.

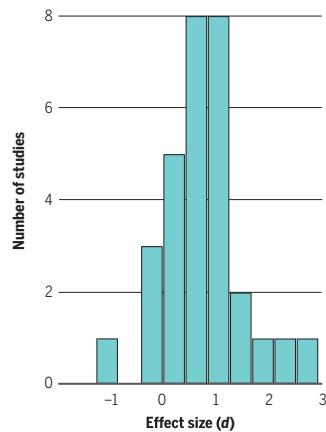
#### Robot behavior

To be effective educational agents, the behavior of social robots must be tailored to support various aspects of learning across different learners and diverse educational contexts. Several studies focused on understanding critical aspects of educational interactions to which robots should respond, as well as determining both what behaviors social robots can use and when to deliver these behaviors to affect learning outcomes.

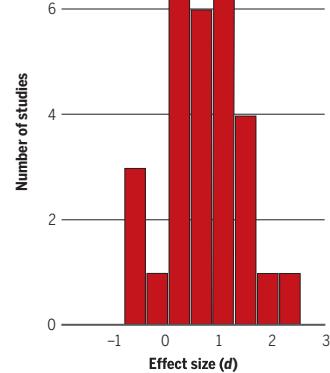
Our meta-review shows that almost any strategy or social behavior of the robot aimed at increasing learning outcomes has a positive effect. We identified the influence of robot behaviors on cognitive outcomes ( $d = 0.69$ ;  $N = 12$ ; 95% CI, 0.56 to 0.83) and affective outcomes ( $d = 0.70$ ;  $N = 32$ ; 95% CI, 0.62 to 0.77).

Similar to findings in the ITS community, robots that personalize what content to provide based on user performance during an interaction can increase cognitive learning gains (43, 44). In addition to the adaptive delivery of learning material, social robots can offer socially supportive behaviors and personalized support for learners within an educational context. Personalized social support, such as using a child's name or referring to previous interactions (45, 46), is the low-hanging fruit of social interaction. More complex prosocial behavior, such as attention-guiding (11), displaying congruent gaze behavior (39), nonverbal immediacy (3), or showing empathy with the learner (36), not only has a positive impact on affective outcomes but also results in increased learning.

#### Cognitive outcomes



#### Affective outcomes



**Fig. 3. Histograms of effect sizes (Cohen's  $d$ ) for all cognitive and affective outcomes of robot tutors in the meta-analysis.** These combine comparisons between robots and alternative educational technologies but also comparisons between different implementations of the robot and its tutoring behavior. In the large majority of results, adding a robot or adding supportive behavior to the robot improves outcomes.

However, just as human tutors must at times sit quietly and allow students the opportunity to concentrate on problem solving, robot tutors must also limit their social behavior at appropriate times based on the cognitive load and engagement of the student (40). The social behavior of the robot must be carefully designed in conjunction with the interaction context and task at hand to enhance the educational interaction and avoid student distraction.

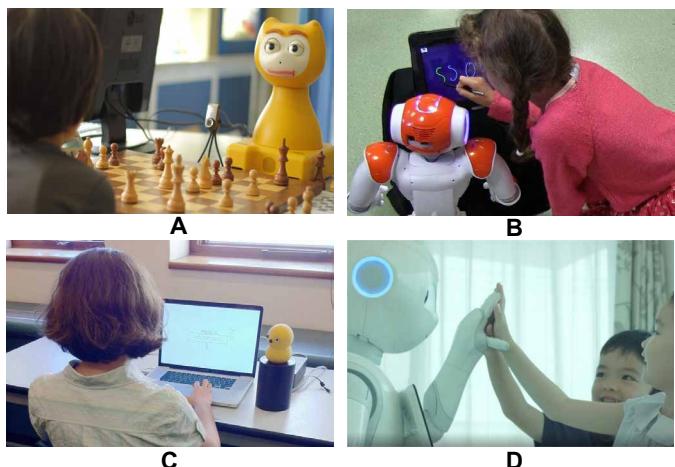
It is possible that the positive cognitive and affective learning outcomes of robot tutors are not directly caused by the robot having a physical presence, but rather the physical presence of the robot promotes social behaviors in the learner that, in turn, foster learning and create a positive learning experience. Robots have been shown to have a positive impact on compliance (26), engagement (21–23), and conformity (20), which, in turn, are conducive to achieving learning gains. Hence, a perhaps valuable research direction is to explore what it is about social robots that affects the first-order outcomes of engagement, persuasion, and compliance.

#### Robot role

Social robots for education include a variety of robots having different roles. Beyond the typical role of a teacher or a tutor, robots can also support learning through peer-to-peer relationships and can support skill consolidation and mastery by acting as a novice. In this section, we provide an overview of the different roles a robot can adopt and what their educational benefits are.

#### Robot as tutor or teacher

As a tutor or teacher, robots provide direct curriculum support through hints, tutorials, and supervision. These types of educational robots, including teaching assistant robots (63), have the longest history of research and development, often targeting curricular domains for young children. Early field studies placed robots into classrooms to observe whether they would have any qualitative impact on the learners' attitude and progress, but current research tends toward controlled experimental trials in both laboratory settings and classrooms (64).



**Fig. 4. Illustrative examples of social robots for learning.** (A) iCat robot teaching young children to play chess (76). (B) Nao robot supporting a child to improve her handwriting (13). (C) Keepon robot tutoring an adult in a puzzle game (27). (D) Pepper robot providing motivation during English classes for Japanese children (74).

A commercial tutor robot called IROBI (Yujin Robotics) was released in the early 2000s. Designed to teach English, IROBI was shown to enhance both concentration on learning activities and academic performance compared with other teaching technology, such as audio material and a web-based application (65).

The focus on younger children links robot education research with other scientific areas, such as language development and developmental psychology (66). On the basis of the earlier work that studied socialization between toddlers and robots in a nursery school (67), a fully autonomous robot was deployed in classrooms. It was shown that the vocabulary skills of 18- to 24-month-old toddlers improved significantly (68). Much of the work in which the robot is used as a tutor focuses on one-to-one interactions, because these offer the greatest potential for personalized education.

In some cases, the robot is used as a novel channel through which a lecture is delivered. In these cases, the robot is not so much interacting with the learners but acts as a teacher or an assistant for the teacher (69). The value of the robot in this case lies in improving attention and motivation in the learners, while the delivery and assessment is done by the human teacher. Here, the delivery is often one to many, with the robot addressing an entire group of learners (33, 63, 69).

#### Robot as peer

Robots can also be peers or learning companions for humans. Not only does a peer have the potential of being less intimidating than a tutor or teacher, peer-to-peer interactions can have significant advantages over tutor-to-student interactions. Robovie was the first fully autonomous robot to be introduced into an elementary school (70). It was an English-speaking robot targeting two grades (first and sixth) of Japanese children. Through field trials conducted over 2 weeks, improvements in English language skills were observed in some children. In one case, longer periods of attention on learning tasks, faster responses, and more accurate responses were shown with a peer robot compared with an identical-looking tutor robot (19). A long-term primary school study showed that a peer-like humanoid robot able to personalize the interaction could increase child learn-

ing of novel subjects (48). Often, the robot is presented as a more knowledgeable peer, guiding the student along a learning trajectory that is neither too easy nor too challenging. However, the role of those robots sometimes becomes ambiguous (tutor versus peer), and it is difficult to place one above the other in general. Learning companions (71), which offer motivational support but otherwise are not tutoring, are also successful cases of a peer-like robot.

#### Robot as novice

Considerable educational benefits can also be obtained from a robot that takes the role of a novice, allowing the student to take on the role of an instructor that typically improves confidence while, at the same time, establishing learning outcomes. This is an instance of learning by teaching, which is widely known in human education, also referred to as the protégé effect (72). This process involves the learner making an effort to teach the robot, which has a direct impact on their own learning outcomes.

The care-receiving robot (CRR) was the first robot designed with the concept of a teachable robot for education (73). A small humanoid robot introduced into English classes improved the vocabulary learning of 3- to 6-year-old Japanese children (5). The robot was designed to make deliberate errors in English vocabulary but could be corrected through instruction by the children. In addition, CRR was shown to engage children more than alternative technology, which eventually led to the release of a commercial product based on the principle of a robot as a novice (74).

This novice role can also be used to teach motor skills. The CoWriter project explored the use of a teachable robot to help children improve their handwriting skills (13). A small humanoid robot in conjunction with a touch tablet helped children who struggled with handwriting to improve their fine motor skills. Here, the children taught the robot, who initially had very poor handwriting, and in the process of doing so, the children reflected on their own writing and showed improved motor skills (13). This suggests that presenting robots as novices has potential to develop meta-cognitive skills in learners, because the learners are committing to instructing the learning material, requiring a higher level of understanding of the material and an understanding of the internal representations of their robot partner.

In our meta-analysis, the robot was predominantly used as a tutor (48%), followed by a role as teacher (38%). In only 9% of studies was the robot presented as a peer or novice (Fig. 1B). The robot was often used to offer one-to-one interactions (65%), with the robot used in a one-to-many teaching scenario in only 30% of the studies (Fig. 1C). In 5%, the robot had mixed interactions, whereby, for example, it first taught more than one student and then had one-on-one interactions during a quiz.

## DISCUSSION

Although an increasing number of studies confirm the promise of social robots for education and tutoring, this Review also lays bare a number of challenges for the field. Robots for learning, and social robotics in general, require a tightly integrated endeavor. Introducing these technologies into educational practice involves solving technical challenges and changing educational practice.

With regard to the technical challenges, building a fluent and contingent interaction between social robots and learners requires the seamless integration of a range of processes in artificial intelligence and robotics. Starting with the input to the system, the robot needs a sufficiently correct interpretation of the social environment

for it to respond appropriately. This requires significant progress in constituent technical fields, such as speech recognition and visual social signal processing, before the robot can access the social environment. Speech recognition, for example, is still insufficiently robust to allow the robot to understand spoken utterances from young children. Although these shortcomings can be resolved by using alternative input media, such as touch screens, this does place a considerable constraint on the natural flow of the interaction. For robots to be autonomous, they must make decisions about which actions to take to scaffold learning. Action selection is a challenging domain at best and becomes more difficult when dealing with a pedagogical environment, because the robot must have an understanding of the learner's ability and progress to allow it to choose appropriate actions. Finally, the generation of verbal and nonverbal output remains a challenge, with the orchestrated timing of verbal and nonverbal actions a prime example. In summary, social interaction requires the seamless functioning of a wide range of cognitive mechanisms. Building artificial social interaction requires the artificial equivalent of these cognitive mechanisms and their interfaces, which is why artificial social interaction is perhaps one of the most formidable challenges in artificial intelligence and robotics.

Introducing social robots in the school curriculum also poses a logistical challenge. The generation of content for social robots for learning is nontrivial, requiring tailor-made material that is likely to be resource-intensive to produce. Currently, the value of the robot lies in tutoring very specific skills, such as mathematics or handwriting, and it is unlikely that robots can take up the wide range of roles a teacher has, such as pedagogical and carer roles. For the time being, robots are mainly deployed in elementary school settings. Although some studies have shown the efficacy of tutoring adolescents and adults, it is unclear whether the approaches that work well for younger children transfer to tutoring older learners.

Introducing robots might also carry risks. For example, studies of ITS have shown that children often do not make the best use of on-demand support and either rely too much on the help function or avoid using help altogether, both resulting in suboptimal learning. Although strategies have been explored to mitigate this particular problem in robots (4), there might be other problems specific to social robots that still need to be identified and for which solutions will be needed.

Social robots have, in the broadest sense, the potential to become part of the educational infrastructure, just as paper, white boards, and computer tablets have. Next to the functional dimension, robots also offer unique personal and social dimensions. A social robot has the potential to deliver a learning experience tailored to the learner, supporting and challenging students in ways unavailable in current resource-limited educational environments. Robots can free up precious time for human teachers, allowing the teacher to focus on what people still do best: providing a comprehensive, empathic, and rewarding educational experience.

Next to the practical considerations of introducing robots in education, there are also ethical issues. How far do we want the education of our children to be delegated to machines, and social robots in particular? Overall, learners are positive about their experience with robots for learning, but parents and teaching staff adopt a more cautious attitude (75). There is much to gain from using robots, but what do we stand to lose? Might robots lead to an impoverished learning experience where what is technologically possible is prioritized over what is actually needed by the learner?

Notwithstanding, robots show great promise when teaching restricted topics, with effect sizes on cognitive outcomes almost matching those of human tutoring. This is remarkable, because our meta-analysis gathered results from a wide range of countries using different robot types, teaching approaches, and deployment contexts. Although the use of robots in educational settings is limited by technical and logistical challenges for now, the benefits of physical embodiment may lift robots above competing learning technologies, and classrooms of the future will likely feature robots that assist a human teacher.

## REFERENCES AND NOTES

1. N. C. Krämer, G. Bente, Personalizing e-Learning. The social effects of pedagogical agents. *Educ. Psychol. Rev.* **22**, 71–87 (2010).
2. J. A. Kulik, J. D. Fletcher, Effectiveness of intelligent tutoring systems: A meta-analytic review. *Rev. Educ. Res.* **86**, 42–78 (2016).
3. J. Kennedy, P. Baxter, E. Senft, T. Belpaeme, in *Proceedings of the International Conference on Social Robotics* (Springer, 2015), pp. 327–336.
4. A. Ramachandran, A. Litoiu, B. Scassellati, in *Proceedings of the 11th ACM/IEEE Conference on Human-Robot Interaction* (IEEE, 2016), pp. 247–254.
5. F. Tanaka, S. Matsuzoe, Children teach a care-receiving robot to promote their learning: Field experiments in a classroom for vocabulary learning. *J. Hum. Robot Interact.* **1**, 78–95 (2012).
6. I. Leite, C. Martinho, A. Paiva, Social robots for long-term interaction: A survey. *Int. J. Soc. Robot.* **5**, 291–308 (2013).
7. J. Han, *Robot-Aided Learning and r-Learning Services* (INTECH Open Access Publisher, 2010).
8. O. Mubin, C. J. Stevens, S. Shahid, A. Al Mahmud, J.-J. Dong, A review of the applicability of robots in education. *J. Technol. Educ. Learning* **1**, 1–7 (2013).
9. J. Gorham, The relationship between verbal teacher immediacy behaviors and student learning. *Commun. Educ.* **37**, 40–53 (1988).
10. P. L. Witt, L. R. Wheless, M. Allen, A meta-analytical review of the relationship between teacher immediacy and student learning. *Commun. Monogr.* **71**, 184–207 (2004).
11. M. Saerbeck, T. Schut, C. Bartneck, M. D. Janse, Expressive robots in education: Varying the degree of social supportive behavior of a robotic tutor, in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI'10 (ACM, 2010), pp. 1613–1622.
12. V. Girotto, C. Lozano, K. Muldner, W. Burleson, E. Walker, Lessons learned from in-school use of rt@q: A robo-tangible learning environment, in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (ACM, 2016), pp. 919–930.
13. D. Hood, S. Lemaignan, P. Dillenbourg, When children teach a robot to write: An autonomous teachable humanoid which uses simulated handwriting, in *Proceedings of the 10th ACM/IEEE International Conference on Human-Robot Interaction* (ACM, 2015), pp. 83–90.
14. A. Litoiu, B. Scassellati, Robotic coaching of complex physical skills, in *Proceedings of the 10th ACM/IEEE International Conference on Human-Robot Interaction* (ACM, 2015), pp. 211–212.
15. J. Fasola, M. Mataric, A socially assistive robot exercise coach for the elderly. *J. Hum. Robot Interact.* **2**, 3–32 (2013).
16. A. Kulkarni, A. Wang, L. Urbina, A. Steinfeld, B. Dias, in *The Eleventh ACM/IEEE International Conference on Human Robot Interaction* (IEEE Press, 2016), pp. 461–462.
17. B. Scassellati, J. Brower, K. Tsui, S. N. Gilani, M. Malzkuhn, B. Manini, A. Stone, G. Kartheiser, A. Merla, A. Shapiro, D. Traum, L. Petitto, Teaching language to deaf infants with a robot and a virtual human, in *Proceedings of the ACM CHI Conference on Human Factors in Computing Systems*, 21 to 26 April 2018, Montréal, Canada (ACM, 2018).
18. R. A. Richert, M. B. Robb, E. I. Smith, Media as social partners: The social nature of young children's learning from screen media. *Child Dev.* **82**, 82–95 (2011).
19. C. Zaga, M. Lohse, K. P. Truong, V. Evers, The effect of a robot's social character on children's task engagement: Peer versus tutor, in *International Conference on Social Robotics* (Springer, 2015), pp. 704–713.
20. J. Kennedy, P. Baxter, T. Belpaeme, Comparing robot embodiments in a guided discovery learning interaction with children. *Int. J. Soc. Robot.* **7**, 293–308 (2015).
21. C. D. Kidd, C. Breazeal, Effect of a robot on user perceptions, in *Proceedings of the 2004 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2004 (IROS 2004) (IEEE, 2004), vol. 4, pp. 3559–3564.
22. J. Wainer, D. J. Feil-Seifer, D. A. Shell, M. J. Mataric, in *Proceedings of the 16th IEEE International Symposium on Robot and Human interactive Communication*, RO-MAN (IEEE, 2007), pp. 872–877.
23. H. Köse, P. Uluer, N. Akalin, R. Yorgancı, A. Özkul, G. Ince, The effect of embodiment in sign language tutoring with assistive humanoid robots. *Int. J. Soc. Robot.* **7**, 537–548 (2015).

24. A. Powers, S. Kiesler, S. Fussell, C. Torrey, Comparing a computer agent with a humanoid robot, in *Proceedings of the 2nd ACM/IEEE International Conference on Human-Robot Interaction* (IEEE, 2007), pp. 145–152.
25. J. Li, The benefit of being physically present: A survey of experimental works comparing copresent robots, telepresent robots and virtual agents. *Int. J. Hum. Comput. Stud.* **77**, 23–37 (2012).
26. W. A. Bainbridge, J. W. Hart, E. S. Kim, B. Scassellati, The benefits of interactions with physically present robots over video-displayed agents. *Int. J. Soc. Robot.* **3**, 41–52 (2011).
27. D. Leyzberg, S. Spaulding, M. Toneva, B. Scassellati, The physical presence of a robot tutor increases cognitive learning gains, in *Proceedings of the 34th Annual Conference of the Cognitive Science Society, CogSci 2012* (2012), pp. 1882–1887.
28. C. D. Kidd, C. Breazeal, A robotic weight loss coach, in *Proceedings of the National Conference on Artificial Intelligence* (MIT Press, 2007), vol. 22, pp. 1985–1986.
29. J. Kennedy, S. Lemaignan, C. Montassier, P. Lavalade, B. Irfan, F. Papadopoulos, E. Senft, T. Belpaeme, Child speech recognition in human-robot interaction: Evaluations and recommendations, in *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction* (ACM/IEEE, 2017), pp. 82–90.
30. P. Baxter, R. Wood, T. Belpaeme, A touchscreen-based ‘sandtray’ to facilitate, mediate and contextualise human-robot social interaction, in *Proceedings of the 7th Annual ACM/IEEE International Conference on Human-Robot Interaction* (ACM, 2012), pp. 105–106.
31. A. Ramachandran, C.-M. Huang, B. Scassellati, Give me a break! Personalized timing strategies to promote learning in robot-child tutoring, in *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction* (ACM, 2017), pp. 146–155.
32. D. Szafir, B. Mutlu, Pay attention! Designing adaptive agents that monitor and improve user engagement, in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI’12* (ACM, 2012), pp. 11–20.
33. I. Leite, M. McCoy, D. Ullman, N. Salomons, B. Scassellati, Comparing models of disengagement in individual and group interactions, in *Proceedings of the 10th Annual ACM/IEEE International Conference on Human-Robot Interaction* (ACM, 2015), pp. 99–105.
34. A. Coninx, P. Baxter, E. Oleari, S. Bellini, B. Bierman, O. B. Henkemans, L. Cañamero, P. Cosi, V. Enescu, R. Ros Espinoza, A. Hiolle, R. Humbert, B. Kiefer, Towards long-term social child-robot interaction: Using multi-activity switching to engage young users. *J. Hum. Robot Interact.* **5**, 32–67 (2016).
35. S. Lemaignan, F. Garcia, A. Jacq, P. Dillenbourg, From real-time attention assessment to “with-me-ness” in human-robot interaction, in *Proceedings of the 11th ACM/IEEE International Conference on Human-Robot Interaction* (IEEE, 2017).
36. I. Leite, G. Castellano, A. Pereira, C. Martinho, A. Paiva, Empathic robots for long-term interaction. *Int. J. Soc. Robot.* **6**, 329–341 (2014).
37. A. Ramachandran, C.-M. Huang, E. Gartland, B. Scassellati, Thinking aloud with a tutoring robot to enhance learning, in *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction* (ACM, 2018), pp. 59–68.
38. C.-M. Huang, B. Mutlu, Modeling and evaluating narrative gestures for humanlike robots, in *Proceedings of the Robotics: Science and Systems Conference, RSS’13* (2013).
39. C.-M. Huang, B. Mutlu, The repertoire of robot behavior: Enabling robots to achieve interaction goals through social behavior. *J. Hum. Robot Interact.* **2**, 80–102 (2013).
40. J. Kennedy, P. Baxter, T. Belpaeme, The robot who tried too hard: Social behaviour of a robot tutor can negatively affect child learning, in *Proceedings of the 10th ACM/IEEE International Conference on Human-Robot Interaction* (ACM, 2015), pp. 67–74.
41. E. Yadollahi, W. Johal, A. Paiva, P. Dillenbourg, When deictic gestures in a robot can harm child-robot collaboration, in *Proceedings of the 17th ACM Conference on Interaction Design and Children* (ACM, 2018), pp. 195–206.
42. G. Gordon, C. Breazeal, Bayesian active learning-based robot tutor for children’s word-reading skills, in *Proceedings of the 29th AAAI Conference on Artificial Intelligence, AAAI-15* (2015).
43. D. Leyzberg, S. Spaulding, B. Scassellati, Personalizing robot tutors to individual learning differences, in *Proceedings of the 9th ACM/IEEE International Conference on Human-Robot Interaction* (ACM, 2014).
44. T. Schodde, K. Bergmann, S. Kopp, Adaptive robot language tutoring based on Bayesian knowledge tracing and predictive decision-making, in *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction* (ACM, 2017), pp. 128–136.
45. J. Janssen, C. van der Wal, M. Neerincx, R. Looije, Motivating children to learn arithmetic with an adaptive robot game, in *Proceedings of the Third International conference on Social Robotics* (ACM, 2011), pp. 153–162.
46. O. A. Blanson Henkemans, B. P. Bierman, J. Janssen, M. A. Neerincx, R. Looije, H. van der Bosch, J. A. van der Giessen, Using a robot to personalise health education for children with diabetes type 1: A pilot study. *Patient Educ. Couns.* **92**, 174–181 (2013).
47. G. Gordon, S. Spaulding, J. K. Westlund, J. J. Lee, L. Plummer, M. Martinez, M. Das, C. Breazeal, Affective personalization of a social robot tutor for children’s second language skills, in *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence* (AAAI, 2016), pp. 3951–3957.
48. P. Baxter, E. Ashurst, R. Read, J. Kennedy, T. Belpaeme, Robot education peers in a situated primary school study: Personalisation promotes child learning. *PLOS ONE* **12**, e0178126 (2017).
49. D. Leyzberg, E. Avrunin, J. Liu, B. Scassellati, Robots that express emotion elicit better human teaching, in *Proceedings of the 6th International Conference on Human-Robot Interaction* (ACM, 2011), pp. 347–354.
50. C. D. Kidd, “Designing for long-term human-robot interaction and application to weight loss,” thesis, Massachusetts Institute of Technology (2008).
51. The meta-analysis data are available at <https://tinyurl.com/ybuyz5vn>.
52. B. Bloom, M. Engelhart, E. Furst, W. Hill, D. Krathwohl, *Taxonomy of Educational Objectives: The Classification of Educational Goals. Handbook I: Cognitive Domain* (Donald McKay, 1956).
53. D. Krathwohl, B. Bloom, B. Masia, *Taxonomy of Educational Objectives: The Classification of Educational Goals. Handbook II: The Affective Domain* (Donald McKay, 1964).
54. D. R. Krathwohl, A revision of bloom’s Taxonomy: An overview. *Theory Pract.* **41**, 212–218 (2002).
55. <https://www.campbellcollaboration.org/escalc/html/EffectSizeCalculator-Home.php>
56. M. W. Lipsey, D. B. Wilson, *Practical Meta-Analysis* (Sage Publications, Inc, 2001).
57. C.-M. Huang, B. Mutlu, Learning-based modeling of multimodal behaviors for humanlike robots, in *Proceedings of the 2014 ACM/IEEE International Conference on Human-Robot Interaction* (ACM, 2014), pp. 57–64.
58. B. S. Bloom, The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educ. Res.* **13**, 4–16 (1984).
59. K. VanLehn, The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educ. Psychol.* **46**, 197–221 (2011).
60. T. W. Draper, W. W. Clayton, Using a personal robot to teach young children. *J. Genet. Psychol.* **153**, 269–273 (1992).
61. M. Imai, T. Ono, H. Ishiguro, Physical relation and expression: Joint attention for human-robot interaction. *IEEE Trans. Ind. Electron.* **50**, 636–643 (2003).
62. B. Mutlu, J. Forlizzi, J. Hodgins, A storytelling robot: Modeling and evaluation of human-like gaze behavior, in *Humanoid Robots, 2006 6th IEEE-RAS International Conference* (IEEE, 2006), pp. 518–523.
63. Z.-J. You, C.-Y. Shen, C.-W. Chang, B.-J. Liu, G.-D. Chen, A robot as a teaching assistant in an English class, in *Proceedings of the Sixth International Conference on Advanced Learning Technologies* (IEEE, 2006), pp. 87–91.
64. T. Belpaeme, P. Vogt, R. Van den Berghe, K. Bergmann, T. Göksun, M. De Haas, J. Kanero, J. Kennedy, A. C. Küntay, O. Oudgenoeg-Paz, F. Papadopoulos, Guidelines for designing social robots as second language tutors. *Int. J. Soc. Robot.* **10**, 1–17 (2018).
65. J.-H. Han, M.-H. Jo, V. Jones, J.-H. Jo, Comparative study on the educational use of home robots for children. *J. Inform. Proc. Syst.* **4**, 159–168 (2008).
66. J. Movellan, F. Tanaka, I. Fasel, C. Taylor, P. Ruvolo, M. Eckhardt, The RUBI project: A progress report, in *Proceedings of the Second ACM/IEEE International Conference on Human-Robot Interaction* (ACM, 2007).
67. F. Tanaka, A. Cicourel, J. R. Movellan, Socialization between toddlers and robots at an early childhood education center. *Proc. Natl. Acad. Sci. U.S.A.* **104**, 17954–17958 (2007).
68. J. R. Movellan, M. Eckhardt, M. Virnes, A. Rodriguez, Sociable robot improves toddler vocabulary skills, in *Proceedings of the 4th ACM/IEEE International Conference on Human-Robot Interaction* (ACM, 2009), pp. 307–308.
69. M. Aleji, A. Meghdari, M. Ghazisaedy, Employing humanoid robots for teaching English language in Iranian junior high-schools. *Int. J. Humanoid Robot.* **11**, 1450022 (2014).
70. T. Kanda, T. Hirano, D. Eaton, H. Ishiguro, Interactive robots as social partners and peer tutors for children: A field trial. *Hum. Comput. Interact.* **19**, 61–64 (2004).
71. N. Lubold, E. Walker, H. Pon-Barry, Effects of voice-adaptation and social dialogue on perceptions of a robotic learning companion, in *The Eleventh ACM/IEEE International Conference on Human Robot Interaction* (IEEE Press, 2017), pp. 255–262.
72. C. C. Chase, D. B. Chin, M. A. Oppezzo, D. L. Schwartz, Teachable agents and the protégé effect: Increasing the effort towards learning. *J. Sci. Educ. Technol.* **18**, 334–352 (2009).
73. F. Tanaka, T. Kimura, The use of robots in early education: A scenario based on ethical consideration, in *Proceedings of the 18th IEEE International Symposium on Robot and Human Interactive Communication* (IEEE, 2009), pp. 558–560.
74. F. Tanaka, K. Isshiki, F. Takahashi, M. Uekusa, R. Sei, K. Hayashi, Pepper learns together with children: Development of an educational application, in *IEEE-RAS 15th International Conference on Humanoid Robots, HUMANOIDS 2015* (IEEE, 2015), pp. 270–275.
75. J. Kennedy, S. Lemaignan, T. Belpaeme, The cautious attitude of teachers towards social robots in schools, in *Proceedings of the Robots 4 Learning Workshop at RO-MAN 2016* (2016).
76. I. Leite, A. Pereira, G. Castellano, S. Mascarenhas, C. Martinho, A. Paiva, Social robots in learning environments: A case study of an empathic chess companion, in *Proceedings of the International Workshop on Personalization Approaches in Learning Environments* (2011).
77. R. R. Hake, Interactive-engagement vs. traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses. *Am. J. Phys.* **66**, 64–74 (1998).
78. C. Barnecker, D. Kulić, E. Croft, S. Zoghbi, Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *Int. J. Soc. Robot.* **1**, 71–81 (2009).

79. R. F. Ferguson, *The Tripod Project Framework* (Tripod, 2008).
80. M. Alemi, A. Meghdari, M. Ghazisaedy, The impact of social robotics on L2 learners' anxiety and attitude in English vocabulary acquisition. *Int. J. Soc. Robot.* **7**, 523–535 (2015).
81. M. Fridin, Storytelling by a kindergarten social assistive robot: A tool for constructive learning in preschool education. *Comput. Educ.* **70**, 53–64 (2014).
82. D. Jolliffe, D. P. Farrington, Development and validation of the basic empathy scale. *J. Adolesc.* **29**, 589–611 (2006).

**Acknowledgments:** We are grateful to E. Ashurst for support in collecting the data for the meta-analysis. **Funding:** This work is partially funded by the H2020 L2TOR project (688014), Japan Society for the Promotion of Science KAKENHI (15H01708), and NSF award 1139078.

**Author contributions:** All authors contributed equally to the manuscript; T.B. and J.K. contributed to the meta-analysis. **Competing interests:** J.K. is a research scientist at Disney Research. **Data and materials availability:** The meta-analysis data are available at <https://tinyurl.com/ybuyz5vn>.

Submitted 31 March 2018

Accepted 23 July 2018

Published 15 August 2018

10.1126/scirobotics.aat5954

**Citation:** T. Belpaeme, J. Kennedy, A. Ramachandran, B. Scassellati, F. Tanaka, Social robots for education: A review. *Sci. Robot.* **3**, eaat5954 (2018).

## Social robots for education: A review

Tony Belpaeme James Kennedy Aditi Ramachandran Brian Scassellati Fumihide Tanaka

*Sci. Robot.*, 3 (21), eaat5954. • DOI: 10.1126/scirobotics.aat5954

### View the article online

<https://www.science.org/doi/10.1126/scirobotics.aat5954>

### Permissions

<https://www.science.org/help/reprints-and-permissions>



## Full length article

# The media inequality: Comparing the initial human-human and human-AI social interactions



Yi Mou, Ph.D. <sup>a,\*</sup>, Kun Xu <sup>b</sup>

<sup>a</sup> Shanghai Jiao Tong University, School of Media & Design, 800 Dongchuan Road, Minhang District, Shanghai, 200240, China

<sup>b</sup> Temple University, School of Media and Communication, 2020 N. 13th Street, Philadelphia, PA, 19122, USA

## ARTICLE INFO

### Article history:

Received 6 November 2016

Received in revised form

20 January 2017

Accepted 28 February 2017

Available online 3 March 2017

### Keywords:

Human-machine communication

The Computers Are Social Actors Paradigm

The cognitive-affective processing system

Artificial intelligence

Chatbot

Social interaction

## ABSTRACT

As human-machine communication has yet to become prevalent, the rules of interactions between human and intelligent machines need to be explored. This study aims to investigate a specific question: During human users' initial interactions with artificial intelligence, would they reveal their personality traits and communicative attributes differently from human-human interactions? A sample of 245 participants was recruited to view six targets' twelve conversation transcripts on a social media platform: Half with a chatbot Microsoft's Little Ice, and half with human friends. The findings suggested that when the targets interacted with Little Ice, they demonstrated different personality traits and communication attributes from interacting with humans. Specifically, users tended to be more open, more agreeable, more extroverted, more conscientious and self-disclosing when interacting with humans than with AI. The findings not only echo Mischel's cognitive-affective processing system model but also complement the Computers Are Social Actors Paradigm. Theoretical implications were discussed.

© 2017 Elsevier Ltd. All rights reserved.

## 1. Introduction

Since Alan Turing proposed the famous question "Can machines think?" in 1950, the emergence of intelligent machines has been witnessed in the past several decades. In the same year, Norbert Wiener envisioned the popularity of interactions between humans and machines and machines and machines (Wiener, 1988). While the idea of machine-machine communication has been testified with 61.5% of traffic on the web being non-human (Klion, 2013), human-machine communication has yet to become prevalent.

Gunkel (2012) proposed that a paradigm shift from computer-mediated communication (CMC) to human-machine communication (HMC) is needed to address the issues associated with communicating with intelligent machines, autonomous decision making systems, and smart devices. Unlike previous generations of machines with few signs of intelligence, those intelligent machines now not only function as a channel of communication process, but also play an active role in participating in communicative interactions.

One of the most representative forms of those intelligent

machines is robots. Robots have been adopted in restaurants, shopping malls, and hospitals. For instance, telepresence robot SAM can work as a nurse assistant to check on senior patients' physical status (Ackerman, 2016). Anderson (2016) observed that the Kirobi Mini robot from Japan could promote people's emotional responses to babies. In addition to the physically embodied robots, online chatbots have also been widely used. Franceschi-Bicchieri (2016) found that a Twitter chatbot in Argentina could trick people into believing its human identity. Today as people have a growing chance of interacting with these digital interlocutors, studying how people initiate and engage in a conversation with machines would lead us to understand our reactions and attitudes towards machines.

Comparing the initial human-machine communication and human-human communication would bring benefits to our interpretation of any potential boundary between humans and machines. Although previous research on media equation has suggested that humans treat media as social actors (Nass & Moon, 2000; Reeves & Nass, 1996), whether media users apply the same level of social responses to media as to humans remains to be explored. The current study on the comparisons between human-human communication and human-machine communication would thus contribute to the theoretical framework of media equation.

\* Corresponding author.

E-mail addresses: [yimou@sjtu.edu.cn](mailto:yimou@sjtu.edu.cn) (Y. Mou), [kun.xu@temple.edu](mailto:kun.xu@temple.edu) (K. Xu).

In addition, engineers and designers could customize their products based on users' personalities in human-machine communication. For example, research has suggested that agreeableness could help reduce interpersonal distance between humans and robots (Takayama & Pantofaru, 2009). In order to improve the user experience of human-machine communication, designers could embed different types of personalities in machines or applications based on users' needs and social responses. Thus, examining the personalities people reveal in human-machine communication would have both theoretical and practical implications.

Overall, this study aims to investigate a specific question: During human users' initial interactions with artificial intelligence (AI), would they reveal their personality traits and communicative attributes differently from human-human interactions? Based on a review of two theoretical frameworks, hypotheses are proposed and tested by an exploratory study.

## 2. Literature review

### 2.1. Human-machine communication

The role machines play in communication process has changed rapidly in recent years. In the past few decades, the scholarship of CMC has viewed machines as a mere channel of information transmission. For instance, a typical question that CMC scholars investigate is whether technological affordances cut off the socio-emotional quality of communication online (Walther, Van Der Heide, Ramirez, Burgoon, & Peña, 2015). However, HMC researchers have interrogated a series of different questions, "What are the boundaries between human and machine? What communicative practices or precepts must be drawn, redrawn or reconsidered to explore these increasingly, or always-already technologized relationships?" (McDowell & Gunkel, 2016, p. 2). As Sundar, Jia, Waddell, and Huang (2015) pointed out, CMC concerns the shortcomings of machine in comparison with face-to-face interaction, whereas HMC is actually attributed to the shortcomings of human mind.

HMC is an ongoing sense-making process between human and machine (Guzman, 2016). In this communication process, how human interlocutors interpret their digital interlocutors and therefore behave accordingly becomes interesting. Will humans treat machines equally as they do to other humans in social interactions, as suggested by the media equation scholars? Or as other researchers posit (e.g., Fischer, Foth, Rohlffing, & Wrede, 2011), they react to machines differently from humans due to the change of interlocutors' nature? We will review the literature from both sides.

### 2.2. The computers are social actors paradigm

Nass and his colleagues were among the pioneers in investigating how humans treat machines in HMC processes. In the 1990s, they proposed the Computers Are Social Actors Paradigm (CASA) (Nass, Fogg, & Moon, 1996; Nass, Steuer, & Tauber, 1994). The CASA paradigm was based on a series of evidence from experiments on human computer interactions. Specifically, Reeves and Nass (1996) selected findings from social science research and replicated the role of humans with computers or televisions. Reeves and Nass (1996) proposed the media equation theory, suggesting that people treat computers and televisions like real people.

Nass and his colleagues have demonstrated that computer users apply politeness to computers (Nass et al., 1994). Nass et al. (1994) found that when a computer asked participants to evaluate its own performance, users tended to have a more positive attitude toward

the computer. However, when a second computer asked participants to evaluate the performance of the first computer they interacted with, users did not show politeness and tended to be more critical of the first computer.

Users also perceive computers to have gender characteristics. Participants reported a computer with a male voice to be more credible and dominant (Nass, Moon, & Green, 1997). A computer with a female voice is seen as having more knowledge about love and relationships, while a computer with a male voice is perceived to be more knowledgeable about technical issues (Nass et al., 1997). Computers' voices can also manifest personalities. Nass and Lee (2001) manipulated the speech rate, volume level, fundamental frequency, and pitch range of computers' voices. They found that introverted participants would like to interact with a computer that has introverted voice, while extroverted participants preferred to talk with an extroverted computer (Nass & Lee, 2001).

This tradition of research is not limited to computers. Nass and Moon (2000) tested users' responses to televisions. They assigned some of the participants to one condition where they were asked to watch two different televisions showing news and entertainment respectively. In the other condition, participants watched news and entertainment on the same television. Nass and Moon (2000) found that the participants in the first condition believed that the two televisions were specialists. They were more informative, serious, and featured better quality. Comparatively, the participants in the second condition believed that the television was a generalist and provided less information and lower quality. Nass and Moon's (2000) research suggests that individuals not only perceive computers as social actors, but also televisions as social actors.

The CASA paradigm not only explained users' perception of machines, but also tested the social interactions between humans and machines. Nass and Moon (2000) found that users applied social norms in conversation with machines. Specifically, users were more likely to disclose private information when computers follow human conversations rules such as gradually shifting from one topic to another. In addition, when computers showed high reciprocity, users were more likely to do self-disclosure (Nass & Moon, 2000).

More recently, the CASA paradigm has been applied to user-chatbot interaction. Edwards, Edwards, Spence, and Shelton (2014) compared Twitter users' perception of chatbots' accounts and human's accounts. Results suggested that users could not differentiate twitter bots from human users. Twitter bots were perceived to be credible, attractive, and efficient in communication as much as humans.

In explaining the CASA paradigm, Nass and Moon (2000) proposed that individuals have not fully evolved to differentiate mediated experience from non-mediated experience. Nass and Moon (2000) argued that individuals are likely to focus on social cues and neglect the asocial characteristics of the entities. These social cues can easily trigger certain social expectations and rules, which lead individuals to use simple scripts that have been applied in the past social interaction.

Although mindlessness is viewed as one of the major reasons for people's social responses to computers, it has been challenged in previous literature. For example, Kanda, Miyashita, Osada, and Ishiguro (2008) found that participants responded to robots' greetings more slowly than to humans' greetings, indicating that participants experienced cognitive activities when responding to robots' behaviors. Fischer et al. (2011) found that participants laughed when responding to robots' greetings, indicating that they detected something unusual during their interaction with robots. These studies suggest that mindlessness may not account for humans' natural and social responses to robots. It is likely that people apply different communication strategies in human-machine communication. In

addition, the work that challenged the explanatory power of mindlessness has centered on users' responses to robots' greetings. Little research has examined whether users' unnatural responses would be found in user-chatbot interaction. Furthermore, few researchers have focused on users' initial interaction with machines. There is a research gap in investigating the factors that lead to people's use of different communication strategies in human-machine communication. Thus in the current study, we attempt to examine users' initial interaction with chatbots from the perspective of their personalities and communication attributes. If users reveal different personalities and communication attributes in human-chatbot communication and human-human communication, researchers could find some potential explanations of users' unnatural responses to machines. The knowledge would also add complements to the CASA paradigm in terms of "when and why mindless behavior will occur" (Nass & Moon, 2000, p. 96).

### 2.3. The cognitive-affective processing system

While the promise of consistence across human-human interactions to human-machine interactions is held by the CASA scholars, social psychologists cast doubts over the psychological invariance that distinctively characterizes an individual across diverse situations (Mischel, 2004). For instance, a student who cheats on a quiz could be highly honest in other situations. Or an individual who are sociable in some events could remain rather shy in others. To solve the so-called "personality paradox", the cognitive-affective processing system (CAPS) was developed by Mischel and colleague (Mischel & Shoda, 1995; Shoda & Mischel, 1998, pp. 175–208).

According to the CAPS model, the personality system contains mental representations consisting of diverse cognitive-affective unites (CAUs). Those CAUs include individuals' core values, beliefs, memories, and so on. They are interconnected and organized, "guided by a stable network of cognitions and effects characteristic for that individual" (Mischel, 2004, p. 11). Individual differences lie in the accessibility of different CAUs. In diverse situations, different CAUs are activated to exhibit behavioral incoherence.

In interpersonal social interactions, individuals tend to obey certain social rules (Burgoon & Jones, 1976). A civil society requires individuals to be polite and sensitive to others' privilege; hence, one has to restrain himself/herself from deviant behaviors. For instance, being afraid of others' moral judgement, individuals will avoid self-disclosing information that contains morality-violating behavior. However, when encountering a nonjudgmental listener such as AI, one's fear of being judged would vanish. In that situation, different CAUs would be activated; he/she might perform boldly and therefore present different personality traits. In a similar vein, one's control over the interactions with a human interlocutor and a machine interlocutor may vary as well. Psychologists have long argued that the control over the events in one's life (including social interactions) demonstrates competence and superiority (Adler, 1930), as an individual constantly matches expectancy against perception in an effort to obtain optimum control (Kelly, 1955). Individuals are not identical in exerting control. For the same person, he/she may barely remain the same control level over all events. When interacting with a machine, some people may feel more confident and take more control over the interacting process while others may feel confused and even intimidated, and consequently implement less control.

It is notable that the CAPS model targets specifically at the inconsistence of personality across situations. Prior research has suggested that personalities could be linked to users' perception of robots (Takayama & Pantofaru, 2009). People with agreeableness would feel less distant from robots when they interact with them,

while people with neuroticism would ask for more personal space when approaching robots (Takayama & Pantofaru, 2009).

On top of the relationship between personalities and interaction with robots, communicative attributes and personality traits are operationally intertwined, as personality has considerable influence on communicative behaviors. For instance, in McCrae and Costa's (1985) big five model of personality (i.e., openness, conscientiousness, agreeableness, extraversion and neuroticism), the level of extraversion is oftentimes reflected by the level of self-disclosure (Hollenbaugh & Ferris, 2014). Therefore, based on the CAPS model, when communicating with a machine, users' self-disclosing behavior should be different from when communicating with a human.

### 2.4. Study overview

In this study, we aim to compare the initial human-AI social interaction with initial human-human social interaction. In particular, we compare the personality traits and communication attributes reflected in human-AI interaction and in human-human interaction. We chose Little Ice, a chatbot developed by Microsoft as an example of AI. The choice was made based on two reasons. First, Microsoft launched Little Ice in China in 2014 and since then it has attracted over 90 millions of users to chat with it (Bingblog, 2016). Little Ice was specifically designed as a 17-year old girl with a lively and outgoing, sometimes naughty personality. This personality contributes to its popularity among users. Second, WeChat, a social network that combines features from Facebook and the mobile application WhatsApp, is currently one of the most popular social media platforms in the world, especially in China. On WeChat, any users can chat with Little Ice as with other human friends. Hence, we could compare one's interactions with Little Ice and with a human friend, while eliminating confounding factors associated with different platform use.

Based on the review of the abovementioned competing theories, we would argue that when conversing with Little Ice, human interlocutors would not remain mindless due to the novelty experience. The presumption of media equation theory is not fulfilled in this context. Hence, the personality traits and communication attributes exhibited in the HMC would be distinct from those in interpersonal communication, as predicted by the CAPS framework. In addition, the situation of introverted self-evaluation also sharply differs from HMC context. According to CAPS, it would be logical to expect differences between human users' self-rated traits and the traits exhibited in HMC.

Altogether, the following hypotheses are postulated:

**H1a.** The personality traits presented by human users in the initial social interaction with a chatbot are different from those presented in the initial social interaction with another human.

**H1b.** The personality traits presented by human users in the initial social interaction with a chatbot are different from the human users' self-rated personality traits.

**H2a.** Human users' self-disclosure level in the initial social interaction with a chatbot is different from that in the initial social interaction with another human.

**H2b.** The level of self-disclosure presented by human users in the initial social interaction with a chatbot is different from the human users' self-rated self-disclosure level.

**H3a.** Human users' level of control over the initial social interaction with a chatbot is different from that in the initial social interaction with another human.

**H3b.** The level of control over the initial social interaction with a chatbot presented by human users is different from the human users' self-rated level of control over social interactions.

### 3. Method

#### 3.1. Procedure

Ten volunteers were recruited through snowball sampling to provide two copies of his/her conversation transcripts on WeChat: One with Little Ice and one with a normal human friend. Those ten volunteers are currently active WeChat users. By the time the conversation transcripts were collected, they had already initiated a chat with Little Ice and their friends on their own will. In other words, when they conversed with Little Ice and human friends, they had no idea of this study at all; so the conversations happened in natural settings. After removing four volunteers' conversation transcripts due to the length or incompleteness of a round of conversation, six volunteers' twelve conversation transcripts were used as the materials for later procedure. Notably, we particularly asked the volunteers to select a conversation transcript with a regular friend, not with a close friend or significant other. Moreover, only the transcripts of their first conversation with Little Ice and friend were retrieved to exclude the variance caused by relationship development. Since those volunteers are the targets of later analysis, we call them "targets" afterwards. Those targets were evaluated by 277 viewers on their personality and communication attributes. Meanwhile, each target filled out measures of their own personality traits and communication attributes.

#### 3.2. Material and stimulus

Based on the targets' original conversation transcripts, three research assistants generated twelve copies of mock-up conversation transcripts after removing sensitive or identifiable information using the Photoshop software. In the name of privacy protection, all names and profile pictures were blocked (see an example of the transcript in Fig. 1).



Fig. 1. An example of the mock-up chatting screenshot of WeChat.

Among those six targets, three are males and three are females. Their age ranges from 19 to 35 years old. While four of them are college students, one works in a private corporate and one works in a major newspaper institute (see their profile detail in Table 1). Their conversations cover a variety of topics based on their identity and interest. For instance, a college female student talks about celebrities and selective courses at school. Another target, an entrepreneur, introduces his corporate. Those conversations seem natural and improvisational. By the time the targets conversed with their human friends, they either just met in virtual groups or were introduced by others in a professional setting or school environment. The conversation topics were included in Table 1.

#### 3.3. Sample

Two hundred and seventy-seven participants (viewers) were recruited in a large public university in Eastern China to read the conversation transcripts and evaluate the targets. After removing sixteen who failed to provide a complete evaluation and another sixteen who were international students, 245 participants' responses were collected for later analysis. To ensure the same baseline, each of the participants was asked to read one of the six targets' two copies of conversation transcripts — one with Little Ice, one with a human friend — based on which they evaluated the targets' personality traits and communicative attributes. When the participants read the transcripts, they had no clue of the purpose of this study. Instead, they were told it was a study about self-presentation on social media in interpersonal communication contexts. Therefore, they were not aware that those two transcripts were from the same target, and one of the conversations happened between a human and a chatbot. Each target was evaluated by 38–43 randomly assigned participants.

Among the 245 viewers, 42.4% of them were males and the rest were females. Their age ranged from 18 to 44 years old. All of them were WeChat users.

#### 3.4. Measure

The questionnaire was originally designed in English. So it was back-translated into Chinese before the questionnaire was administered. The question items in the targets' self-evaluation questionnaire and the viewers' questionnaire were identical, except the differences in sentence structure. For example, in the self-evaluation questionnaire, the direction was "How much do you agree with the following statements that describe yourself?" In contrast, in the viewers' questionnaire, the direction was "How much do you agree with the following statements that describe the target?"

##### 3.4.1. Personality

McCord's (2002) five-factor personality scale was employed to measure personality traits. Although longer versions of measures were available, the 50-item version was used to control the length of the questionnaire. Ten items measured each of the five traits: openness, conscientiousness, agreeableness, extraversion and neuroticism. Specifically, examples of openness (to new experiences) include "have a vivid imagination" and "enjoy hearing new ideas." Examples of conscientiousness are "make plans and stick to them" and "am always prepared." Examples of agreeableness include "believe that others have good intentions" and "make people feel at ease." Examples of extraversion include "am skilled in handling social situations" and "know how to captivate people." Examples of neuroticism are "am not easily bothered by things (reverse coded)" and "feel comfortable with myself (reverse coded)." All the fifty items were measured on a 7-point Likert scale

**Table 1**

The profile detail of each target.

ID	Sex	Age	Profession	Self-rated personality traits	Self-rated level of self-disclosure	Self-rated level of control	Human friend and chatting topic
1	M	35	Entrepreneur	O = 5.60 N = 2.50 A = 5.60 E = 4.60 C = 5.20	6.00	4.00	A female introduced by a professional connection; introducing his corporate.
2	M	25	Journalist	O = 6.00 N = 4.20 A = 3.20 E = 4.10 C = 4.30	4.20	4.50	An unfamiliar female coworker; small talks associated with working setting.
3	M	24	College student	O = 4.30 N = 3.20 A = 4.60 E = 4.40 C = 4.30	4.90	3.83	Cannot tell the gender; small talks associated with the first impressions for each other.
4	F	19	College student	O = 5.60 N = 2.60 A = 5.40 E = 3.90 C = 5.00	5.40	4.00	Cannot tell the gender; on how to select courses on school's system.
5	F	19	College student	O = 5.17 N = 3.17 A = 5.00 E = 4.00 C = 4.00	5.12	4.30	A female student; on student union work.
6	F	20	College student	O = 4.90 N = 3.70 A = 4.80 E = 3.30 C = 5.10	4.44	5.17	A female student; on working on a project together.

from (1) *Strongly Disagree* to (7) *Strongly Agree*. The robustness of this scale has been testified by Cooper, Golden and Socha (2013). The reliability Cronbach's alphas for each factor were 0.78, 0.64, 0.61, 0.78, and 0.70 in that order.

#### 3.4.2. Self-disclosure

The level of self-disclosure was gauged by the scale developed by Miller, Berg, and Archer (1983). The participants were asked to indicate the degree to which they agree with the statements such as "People frequently tell me/this target about themselves." All the ten items were measured on a 7-point Likert scale from (1) *Strongly Disagree* to (7) *Strongly Agree*. The reliability Cronbach's alpha was 0.92.

#### 3.4.3. Control over social interactions

The level of control over social interactions was measured by Shulman, Laursen, Kalman, and Karpovsky (1997) scale after removing two inapplicable items. The participants were asked to indicate the degree to which they agree with the statements such as "This target prefers/You prefer that everyone acts according to his/her/your decisions." All the six items were measured on a 7-point Likert scale from (1) *Strongly Disagree* to (7) *Strongly Agree*. The reliability Cronbach's alpha was 0.70.

#### 3.4.4. Social media use

Targets' and viewers' social media use was gauged by asking them to indicate how much time they spend on various social media platforms each day (including social media on mobile devices) such as WeChat and microblogs. The frequency of social media use was measured on a 7-point scale from (1) *never or barely* to (7) *more than six times each day*. To gauge their use proficiency of WeChat, the participants were also asked to estimate the range of the number of their friends on WeChat, from (1) *less than 50* to (7)

*more than 500*.

#### 3.4.5. Demographics

The targets' and viewers' sex and age were also measured in the questionnaire. In particular, after reading each conversation transcript, the viewers were asked to guess the sex of the target. The responses were coded into (1) true or (0) false.

#### 3.5. Manipulation check

Due to the use of the cover story, no real manipulation check was conducted. But based on the reaction of the viewers, they appeared not to have any doubt over the cover story. The most powerful demonstration of manipulation check was the results, which indicated a series of significant differences of ratings based on two different conversation transcripts (see below). Therefore, we concluded that the manipulation was successful.

### 4. Results

As the viewers spent more than 5 h ( $M = 5.21$ ,  $SD = 2.50$ ) online daily on average, they demonstrated a heavy use of the social media. They checked WeChat multiple times, and spent an average of 2.44 h ( $SD = 2.27$ ) on WeChat each day. The average number of WeChat friends was over 300 ( $M = 4.22$ ,  $SD = 1.59$ ). The second most used social media platform was microblogging service, as the average daily use time was 0.87 h ( $SD = 1.17$ ).

The means and standard deviations of self-rated levels of personality traits and communication attributes of each target were reported in Table 1.

H1-3 predicted significant differences (a) between the personality traits and communication attributes presented by human users in the initial social interaction with a chatbot and the initial social interaction with another human and (b) between those presented by human users in the initial social interaction with a chatbot and their self-rated ones. A series of paired *t*-test analyses were conducted (see Table 2). Notably, in testing H1-3, we combined the evaluations on those six targets together to save space. For the comparisons on each target, please see Appendix.

For the personality trait of openness, the self-rated level was the highest ( $M = 5.27$ ,  $SD = 0.54$ ), followed by the level rated on human-human interaction ( $M = 4.10$ ,  $SD = 0.66$ ). The level rated on human-AI interaction was the lowest ( $M = 3.87$ ,  $SD = 0.52$ ). There existed significant differences among those three evaluations:  $t_{\text{self-AI}}(244) = 29.67$ ,  $p < 0.001$ ;  $t_{\text{self-human}}(244) = 20.28$ ,  $p < 0.001$ ; and  $t_{\text{AI-human}}(244) = -4.27$ ,  $p < 0.001$ . The trait of agreeableness followed the same pattern: the self-rated level was the highest ( $M = 4.79$ ,  $SD = 0.77$ ), followed by the level rated on human-human interaction ( $M = 4.37$ ,  $SD = 0.58$ ). The level rated on human-AI interaction was the lowest ( $M = 3.84$ ,  $SD = 0.60$ ). Those three evaluations were significant from each other:  $t_{\text{self-AI}}(244) = 16.85$ ,  $p < 0.001$ ;  $t_{\text{self-human}}(244) = 6.29$ ,  $p < 0.001$ ; and  $t_{\text{AI-human}}(244) = -9.09$ ,  $p < 0.001$ . The trait of conscientiousness fell into the same category as well. The self-rated level was the highest ( $M = 4.66$ ,  $SD = 0.47$ ), followed by the level rated on human-human interaction ( $M = 4.29$ ,  $SD = 0.59$ ). The level rated on human-AI interaction was the lowest ( $M = 3.80$ ,  $SD = 0.53$ ). Those three evaluations were significant from each other:  $t_{\text{self-AI}}(244) = 18.93$ ,  $p < 0.001$ ;  $t_{\text{self-human}}(244) = 7.25$ ,  $p < 0.001$ ; and  $t_{\text{AI-human}}(244) = -8.98$ ,  $p < 0.001$ .

Interestingly, the trait of neuroticism yielded an opposite pattern: the level rated on human-AI interaction was the highest ( $M = 3.98$ ,  $SD = 0.56$ ), followed by the level rated on human-human interaction ( $M = 3.69$ ,  $SD = 0.61$ ) and self-rated level ( $M = 3.21$ ,  $SD = 0.59$ ). Those three evaluations were significant from each

other as well:  $t_{\text{self-AI}}(244) = -14.62$ ,  $p < 0.001$ ;  $t_{\text{self-human}}(244) = -9.56$ ,  $p < 0.001$ ; and  $t_{\text{AI-human}}(244) = 5.36$ ,  $p < 0.001$ . The trait of extraversion reflected a different pattern: the level rated on human-human interaction was the highest ( $M = 4.21$ ,  $SD = 0.73$ ), followed by self-rated level ( $M = 4.05$ ,  $SD = 0.41$ ); while the level rated on human-AI interaction was the lowest ( $M = 3.79$ ,  $SD = 0.76$ ). Those three evaluations were significant from each other as well:  $t_{\text{self-AI}}(244) = 5.04$ ,  $p < 0.001$ ;  $t_{\text{self-human}}(244) = -3.47$ ,  $p < 0.001$ ; and  $t_{\text{AI-human}}(260) = -5.74$ ,  $p < 0.001$ . Therefore, H1a and H1b were supported.

As for the self-disclosure level, the self-rated level was the highest ( $M = 5.01$ ,  $SD = 0.66$ ), followed by the level rated on human-human interaction ( $M = 4.03$ ,  $SD = 0.98$ ) and the level rated on human-AI interaction ( $M = 3.30$ ,  $SD = 0.84$ ). Significant differences existed between those three levels:  $t_{\text{self-AI}}(244) = 24.53$ ,  $p < 0.001$ ;  $t_{\text{self-human}}(244) = 12.70$ ,  $p < 0.001$ ; and  $t_{\text{AI-human}}(260) = -7.67$ ,  $p < 0.001$ . Hence, H2a and H2b were supported.

As for the level of control, the self-rated level was the highest ( $M = 4.30$ ,  $SD = 0.49$ ), followed by the level rated on human-human interaction ( $M = 4.10$ ,  $SD = 0.81$ ) and the level rated on human-AI interaction ( $M = 4.01$ ,  $SD = 0.69$ ). Significant differences existed between the self-rated level and other two levels:  $t_{\text{self-AI}}(244) = 3.61$ ,  $p < 0.001$ ;  $t_{\text{self-human}}(244) = 3.95$ ,  $p < 0.001$ . But there was no significant difference between the latter two levels:  $t_{\text{AI-human}}(244) = 1.31$ , n.s. Therefore, H3a was not supported; but H3b was supported.

## 5. Discussion

This study set out to detect the discrepancy between the initial social interaction between human and AI and that between humans. The findings suggested that when WeChat users interacted with Little Ice, they demonstrated different personality traits from interactions with humans. Specifically, users tended to be more open, more agreeable, more extroverted, more conscientious and self-disclosing when interacting with humans than with AI. In contrast, they also showed higher level of neuroticism with AI than with humans. In sum, human users demonstrated more socially desirable traits in communicating with humans than with AI. The only exception was the level of control over social interactions, as

**Table 2**  
Three types of evaluations by viewers on personality and communicative attributes.

Trait	Mean	SD	Comparison (all df = 244)
O <sub>self</sub>	5.27	0.54	$t_{\text{self-AI}} = 29.67^{***}$
O <sub>w/AI</sub>	3.87	0.52	$t_{\text{self-human}} = 20.28^{***}$
O <sub>w/human</sub>	4.10	0.66	$t_{\text{AI-human}} = -4.27^{***}$
N <sub>self</sub>	3.21	0.59	$t_{\text{self-AI}} = -14.62^{***}$
N <sub>w/AI</sub>	3.98	0.56	$t_{\text{self-human}} = -9.56^{***}$
N <sub>w/human</sub>	3.69	0.61	$t_{\text{AI-human}} = 5.36^{***}$
A <sub>self</sub>	4.79	0.77	$t_{\text{self-AI}} = 16.85^{***}$
A <sub>w/AI</sub>	3.84	0.60	$t_{\text{self-human}} = 6.29^{***}$
A <sub>w/human</sub>	4.37	0.58	$t_{\text{AI-human}} = -9.09^{***}$
E <sub>self</sub>	4.05	0.41	$t_{\text{self-AI}} = 5.04^{***}$
E <sub>w/AI</sub>	3.79	0.76	$t_{\text{self-human}} = -3.47^{**}$
E <sub>w/human</sub>	4.21	0.73	$t_{\text{AI-human}} = -5.74^{***}$
C <sub>self</sub>	4.66	0.47	$t_{\text{self-AI}} = 18.93^{***}$
C <sub>w/AI</sub>	3.80	0.53	$t_{\text{self-human}} = 7.25^{***}$
C <sub>w/human</sub>	4.29	0.59	$t_{\text{AI-human}} = -8.98^{***}$
SD <sub>self</sub>	5.01	0.66	$t_{\text{self-AI}} = 24.53^{***}$
SD <sub>w/AI</sub>	3.30	0.84	$t_{\text{self-human}} = 12.70^{***}$
SD <sub>w/human</sub>	4.03	0.98	$t_{\text{AI-human}} = -7.67^{***}$
Control <sub>self</sub>	4.30	0.49	$t_{\text{self-AI}} = 3.61^{***}$
Control <sub>w/AI</sub>	4.10	0.81	$t_{\text{self-human}} = 3.95^{***}$
Control <sub>w/human</sub>	4.01	0.69	$t_{\text{AI-human}} = 1.31$

Note: O = Openness; C = Conscientiousness; A = Agreeableness; E = Extraversion; N = Neuroticism; SD = Self-disclosure.

\*\*\* $p < 0.001$ , \*\* $p < 0.01$ .

no significant difference was detected between conversing with humans and with AI.

The findings suggest that users apply different strategies to interact with AI from with humans. The results echo Mischel's CAPS model. When individuals encounter different types of interlocutors, various cognitive-affective unites will be activated. The activation further leads human users to present different personalities. As the CAPS provides a general framework to predict and explain the results, we need to delve into the more specific HMC frameworks.

The findings in this study may complement the CASA paradigm. Nass and Moon (2000) argued that users mindlessly apply social scripts from human-human interaction to human-computer interaction. Reeves and Nass (1996) further used evolutionary psychology to argue that computer users have not evolved enough to distinguish mediated environments from non-mediated environments. The finding in the current study may provide a different perspective of the narrative. If users are aware that they will interact with an AI that is supposed to act like real people, users will show less openness and less extraversion. It is consistent with the prior research finding that people who believed that they would interact with a robot would report lower perceived attractiveness than those who believed that they would interact with a person (Spence, Westerman, Edwards, & Edwards, 2014). Meanwhile, the naughty performances of Little Ice led users to react in a more neurotic way. On one hand, the findings corroborate previous studies in that mindlessness may not be explanatory in some contexts (Amalberti, Carbonell, & Falzon, 1993; Fischer et al., 2011; Kanda, Miyashita, Osada, Haikawa, & Ishiguro, 2008). On the other hand, it should be noted that users' mindless responses occur only when technologies show "enough cues to lead the person to categorize it as worthy of social responses" (Nass & Moon, 2000, p. 83). Thus, it is possible that Little Ice only demonstrated the social cues that evoke a certain degree of social responses but not enough to elicit the same level of responses to humans.

Among the five big personalities, users were perceived to have higher neuroticism in communicating with Little Ice. The result may corroborate Nass and Lee's (2001) finding that computers users preferred to interact with those that have similar personalities to them. As Little Ice was designed to be a naughty girl that can tell jokes, recite poetry, tell horror stories, and so on, users may prefer to respond to Little Ice in a more neurotic way. Meanwhile, the amalgam of AI's naughty personality and the multiple social functions might have led users to feel insecure and reluctant to disclose their information to AI.

In addition, Duffy and Zawieska's (2012) analysis of the different conditions where users suspend their disbelief in social robots could be a good reference to the results. Though Little Ice was endowed with different response mechanisms, the degree of bi-directionality and the strangeness in the conversation between humans and AI may determine how much users suspend their disbelief and build up their trust in the AI (Duffy & Zawieska, 2012). Despite the multiple social functions and designs of Little Ice, it is likely that human users can still tell that Little Ice's responses were not as natural as human conversation. Therefore, the low suspension of disbelief may inhibit users from demonstrating their personalities.

Going beyond the debate surrounding the media equation theory or CASA paradigm, the results of this study also shed light on general social relationships with machines. Based on a discourse analysis, Shechtman and Horowitz (2003) found that when participants believed that they were talking to a person instead of a computer, participants used more words and spent more time in conversation. More importantly, participants used statements about relationships (such as "Well, I definitely would be thankful to have you by my side in this situation.") in human-human

interactions four times than in human-computer interactions. Our research is congruent with the Shechtman and Horowitz (2003) study in that users appear to be restrained in conversing with AI.

Lack of goals may also account for the low level of personality demonstration in conversation with chatbots. In human-human interaction, conversation is goal-driven. Three main categories of goals have been identified in prior research: Task goals, communication goals, and relationship goals (Clark, 1996; Hobbs & Evans, 1980). Those goals help set the tone of our daily conversation. But as HMC is an emerging communication phenomenon, humans may not be able to find appropriate motivation to develop social relationships with machines. That could be the reason why conversing with a chatbot brought about lower ratings on personality.

Another finding is the striking difference between the targets' self-evaluation and viewers' evaluation based on their interactions with Little Ice. The targets tended to rate themselves as more socially desirable, i.e., being agreeable and conscientious, but their interactions with AI tells a different story. The difference may lead us to further reflect on which version of the targets is the true self. Is it the person talking with AI or the person talking with his/her friends? The discrepancies between targets' self-evaluation and viewers' evaluation reflects the prior debates on personality as a trait versus a state. Steyer, Schmitt, and Eid (1999) suggested that the concept of personality can be operationalized as both trait and state. Thus, this "You think you are nice, but you're actually mean to AI" or "revealing the true self to AI" narrative could direct researchers to further inquire into the contexts where users' demonstrated personalities as a trait versus as a state.

Several limitations need to be considered in interpreting the results of this study. First, although Microsoft claimed that over 90 million users have conversed with Little Ice, it is difficult to find suitable targets for this study due to the strict criteria of conversation transcript. That is why only six targets were recruited in the study. Hence, the generalizability of this study is limited. Moreover, this study was conducted in China. While Chinese culture emphasizes compliance with social rules, individuals in Chinese culture may feel more pressured to behave in a socially desirable way in interpersonal communication contexts than their counterparts in other cultures (Hofstede, 1984). Stuart (2016) also suggested that humans' reactions to social robots could be affected by their cultural backgrounds. Future research may consider a cross-cultural comparison of individuals' attitude toward chatbot. In addition, the current study did not investigate the impression formation process from the viewers' perspective. In other words, we did not probe into what cues caused the viewers' judgments on the targets' personality traits and communicative attributes. Future study may use the lens model approach (Hall, Pennington, & Lueders, 2014) to explore this question.

As an exploratory study, this project did not control for some potential confounding variables. For instance, we did not control for the gender and age of the human friend, since we did not want the interlocutors to initiate a conversation upon our request. Instead, we collected the conversation transcripts of natural conversations. This choice might suffer from lower internal validity, but the external validity was boosted. Although Microsoft artificially assigned gender, age and personality traits to Little Ice, Little Ice's conversational response is based on the big data from open public online sites. That is why we did not equate Little Ice with a regular 17-year old girl. The personality of the targeted interlocutor would be an underlying confounding factor as well. But the results indicated that each individual's responses remain consistent with the overall pattern (see Appendix).

On the last note, HMC heavily depends on the evolution of technology. Past generations of chatbot such as ELIZA could only

provide scripted conversational responses, while Little Ice responds autonomously based on the big data of the Internet (Bingblog, 2014). Along with the fast development of speech recognition technology and other similar technologies, human-machine interface is becoming more and more natural. Therefore, it would be premature to draw conclusions on how humans socially react to machines purely based on today's technology. Future studies should proceed to study human-machine relationships.

## Acknowledgements

Supported by the Science Foundation of Ministry of Education of China, Grant no. 14YJC860029. The authors would like to thank YIN Zichun, CHENG Kangzhe and HE Chongyu for their assistance in data collection and entry.

## Appendix. The comparisons of each target

For Target 1:

Trait	Mean	SD	Comparison (all df = 41)
O <sub>self</sub>	5.60	—	t <sub>self-AI</sub> = 20.26***
O <sub>w/AI</sub>	3.93	0.53	t <sub>self-human</sub> = 20.22***
O <sub>w/human</sub>	4.12	0.47	t <sub>AI-human</sub> = -1.92#
N <sub>self</sub>	2.50	—	t <sub>self-AI</sub> = -21.95***
N <sub>w/AI</sub>	4.09	0.47	t <sub>self-human</sub> = -16.59***
N <sub>w/human</sub>	3.70	0.47	t <sub>AI-human</sub> = 3.58**
A <sub>self</sub>	5.60	—	t <sub>self-AI</sub> = 17.39***
A <sub>w/AI</sub>	3.93	0.62	t <sub>self-human</sub> = 14.52***
A <sub>w/human</sub>	4.31	0.57	t <sub>AI-human</sub> = -3.02**
E <sub>self</sub>	4.60	—	t <sub>self-AI</sub> = 5.35***
E <sub>w/AI</sub>	3.95	0.79	t <sub>self-human</sub> = 4.29***
E <sub>w/human</sub>	4.21	0.59	t <sub>AI-human</sub> = -1.68#
C <sub>self</sub>	5.20	—	t <sub>self-AI</sub> = 21.03***
C <sub>w/AI</sub>	3.79	0.43	t <sub>self-human</sub> = 10.25***
C <sub>w/human</sub>	4.43	0.48	t <sub>AI-human</sub> = -6.64***
SD <sub>self</sub>	6.00	—	t <sub>self-AI</sub> = 20.31***
SD <sub>w/AI</sub>	3.52	0.83	t <sub>self-human</sub> = 26.75***
SD <sub>w/human</sub>	3.95	0.52	t <sub>AI-human</sub> = -3.40**
Control <sub>self</sub>	4.00	—	t <sub>self-AI</sub> = -0.03
Control <sub>w/AI</sub>	4.00	0.82	t <sub>self-human</sub> = -1.69#
Control <sub>w/human</sub>	4.18	0.68	t <sub>AI-human</sub> = -1.23

\*p < .05; \*\*p < .01; \*\*\*p < .001.

For Target 2:

Trait	Mean	SD	Comparison (all df = 38)
O <sub>self</sub>	6.00	—	t <sub>self-AI</sub> = 29.71***
O <sub>w/AI</sub>	3.83	0.46	t <sub>self-human</sub> = 14.95***
O <sub>w/human</sub>	4.24	0.73	t <sub>AI-human</sub> = -2.85**
N <sub>self</sub>	4.20	—	t <sub>self-AI</sub> = 4.12***
N <sub>w/AI</sub>	3.81	0.59	t <sub>self-human</sub> = 4.66***
N <sub>w/human</sub>	3.76	0.58	t <sub>AI-human</sub> = 0.31
A <sub>self</sub>	3.20	—	t <sub>self-AI</sub> = -5.14***
A <sub>w/AI</sub>	3.53	0.40	t <sub>self-human</sub> = -16.40***
A <sub>w/human</sub>	4.66	0.58	t <sub>AI-human</sub> = -8.55***
E <sub>self</sub>	4.10	—	t <sub>self-AI</sub> = 3.40**
E <sub>w/AI</sub>	3.77	0.61	t <sub>self-human</sub> = -2.49*
E <sub>w/human</sub>	4.42	0.80	t <sub>AI-human</sub> = -3.76**
C <sub>self</sub>	4.30	—	t <sub>self-AI</sub> = 3.79**
C <sub>w/AI</sub>	3.91	0.64	t <sub>self-human</sub> = -0.56
C <sub>w/human</sub>	4.35	0.54	t <sub>AI-human</sub> = -2.77**
SD <sub>self</sub>	4.20	—	t <sub>self-AI</sub> = 7.23***
SD <sub>w/AI</sub>	3.23	0.84	t <sub>self-human</sub> = -0.50
SD <sub>w/human</sub>	4.30	1.26	t <sub>AI-human</sub> = -3.40**
Control <sub>self</sub>	4.50	—	t <sub>self-AI</sub> = 1.17
Control <sub>w/AI</sub>	4.34	0.84	t <sub>self-human</sub> = 5.41***
Control <sub>w/human</sub>	3.99	0.59	t <sub>AI-human</sub> = 2.02#

\*p < .05; \*\*p < .01; \*\*\*p < .001.

## For Target 3:

Trait	Mean	SD	Comparison (all df = 37)
O <sub>self</sub>	4.30	—	t <sub>self-AI</sub> = 4.28***
O <sub>w/AI</sub>	3.98	0.46	t <sub>self-human</sub> = -2.71*
O <sub>w/human</sub>	4.63	0.75	t <sub>AI-human</sub> = -3.98***
N <sub>self</sub>	3.20	—	t <sub>self-AI</sub> = -6.98***
N <sub>w/AI</sub>	3.93	0.64	t <sub>self-human</sub> = -0.98
N <sub>w/human</sub>	3.32	0.75	t <sub>AI-human</sub> = 4.51***
A <sub>self</sub>	4.60	—	t <sub>self-AI</sub> = 5.74***
A <sub>w/AI</sub>	4.00	0.64	t <sub>self-human</sub> = 3.62**
A <sub>w/human</sub>	4.24	0.61	t <sub>AI-human</sub> = -1.83#
E <sub>self</sub>	4.40	—	t <sub>self-AI</sub> = 6.11***
E <sub>w/AI</sub>	3.66	0.75	t <sub>self-human</sub> = -0.83
E <sub>w/human</sub>	4.52	0.90	t <sub>AI-human</sub> = -3.59**
C <sub>self</sub>	4.30	—	t <sub>self-AI</sub> = 5.60***
C <sub>w/AI</sub>	3.92	0.42	t <sub>self-human</sub> = -0.98
C <sub>w/human</sub>	4.41	0.70	t <sub>AI-human</sub> = -3.95***
SD <sub>self</sub>	4.90	—	t <sub>self-AI</sub> = 9.54***
SD <sub>w/AI</sub>	3.26	1.06	t <sub>self-human</sub> = 4.05***
SD <sub>w/human</sub>	4.15	1.14	t <sub>AI-human</sub> = -2.89**
Control <sub>self</sub>	3.83	—	t <sub>self-AI</sub> = -7.21***
Control <sub>w/AI</sub>	4.43	0.51	t <sub>self-human</sub> = -3.21**
Control <sub>w/human</sub>	4.19	0.70	t <sub>AI-human</sub> = 1.59

\*p &lt; .05; \*\*p &lt; .01; \*\*\*p &lt; .001.

## For Target 6:

Trait	Mean	SD	Comparison (all df = 39)
O <sub>self</sub>	4.90	—	t <sub>self-AI</sub> = 17.07***
O <sub>w/AI</sub>	3.66	0.46	t <sub>self-human</sub> = 11.60***
O <sub>w/human</sub>	3.59	0.71	t <sub>AI-human</sub> = 0.47
N <sub>self</sub>	3.70	—	t <sub>self-AI</sub> = -5.94***
N <sub>w/AI</sub>	4.10	0.43	t <sub>self-human</sub> = -7.13***
N <sub>w/human</sub>	4.19	0.43	t <sub>AI-human</sub> = -0.91
A <sub>self</sub>	4.80	—	t <sub>self-AI</sub> = 8.21***
A <sub>w/AI</sub>	4.14	0.51	t <sub>self-human</sub> = 7.42***
A <sub>w/human</sub>	4.15	0.55	t <sub>AI-human</sub> = -0.06
E <sub>self</sub>	3.30	—	t <sub>self-AI</sub> = -0.30
E <sub>w/AI</sub>	3.34	0.85	t <sub>self-human</sub> = -2.58*
E <sub>w/human</sub>	3.62	0.77	t <sub>AI-human</sub> = -1.47
C <sub>self</sub>	5.10	—	t <sub>self-AI</sub> = 17.66***
C <sub>w/AI</sub>	3.63	0.52	t <sub>self-human</sub> = 18.80***
C <sub>w/human</sub>	3.79	0.44	t <sub>AI-human</sub> = -1.41
SD <sub>self</sub>	4.44	—	t <sub>self-AI</sub> = 14.66***
SD <sub>w/AI</sub>	3.08	0.59	t <sub>self-human</sub> = 9.62***
SD <sub>w/human</sub>	3.14	0.85	t <sub>AI-human</sub> = -0.47
Control <sub>self</sub>	5.17	—	t <sub>self-AI</sub> = 8.76***
Control <sub>w/AI</sub>	3.64	1.11	t <sub>self-human</sub> = 8.26***
Control <sub>w/human</sub>	3.94	0.94	t <sub>AI-human</sub> = -1.31

\*p &lt; .05; \*\*p &lt; .01; \*\*\*p &lt; .001.

## For Target 4:

Trait	Mean	SD	Comparison (all df = 42)
O <sub>self</sub>	5.60	—	t <sub>self-AI</sub> = 17.18***
O <sub>w/AI</sub>	4.06	0.59	t <sub>self-human</sub> = 24.84***
O <sub>w/human</sub>	3.92	0.44	t <sub>AI-human</sub> = 0.95
N <sub>self</sub>	2.60	—	t <sub>self-AI</sub> = -17.23***
N <sub>w/AI</sub>	3.76	0.44	t <sub>self-human</sub> = -13.42***
N <sub>w/human</sub>	3.67	0.52	t <sub>AI-human</sub> = 0.82
A <sub>self</sub>	5.40	—	t <sub>self-AI</sub> = 23.02***
A <sub>w/AI</sub>	3.93	0.42	t <sub>self-human</sub> = 10.91***
A <sub>w/human</sub>	4.39	0.61	t <sub>AI-human</sub> = -3.90***
E <sub>self</sub>	3.90	—	t <sub>self-AI</sub> = -0.94
E <sub>w/AI</sub>	4.01	0.79	t <sub>self-human</sub> = -5.08***
E <sub>w/human</sub>	4.30	0.52	t <sub>AI-human</sub> = -1.69#
C <sub>self</sub>	5.00	—	t <sub>self-AI</sub> = 14.36***
C <sub>w/AI</sub>	3.93	0.49	t <sub>self-human</sub> = 7.97***
C <sub>w/human</sub>	4.41	0.49	t <sub>AI-human</sub> = -3.83***
SD <sub>self</sub>	5.40	—	t <sub>self-AI</sub> = 15.69***
SD <sub>w/AI</sub>	3.62	0.74	t <sub>self-human</sub> = 9.79***
SD <sub>w/human</sub>	4.30	0.74	t <sub>AI-human</sub> = -3.20**
Control <sub>self</sub>	4.00	—	t <sub>self-AI</sub> = 3.45**
Control <sub>w/AI</sub>	3.81	0.36	t <sub>self-human</sub> = 0.78
Control <sub>w/human</sub>	3.93	0.59	t <sub>AI-human</sub> = -1.09

\*p &lt; .05; \*\*p &lt; .01; \*\*\*p &lt; .001.

## For Target 5:

Trait	Mean	SD	Comparison (all df = 42)
O <sub>self</sub>	5.17	—	t <sub>self-AI</sub> = 17.81***
O <sub>w/AI</sub>	3.74	0.53	t <sub>self-human</sub> = 22.34***
O <sub>w/human</sub>	4.16	0.30	t <sub>AI-human</sub> = -3.77**
N <sub>self</sub>	3.17	—	t <sub>self-AI</sub> = -10.32***
N <sub>w/AI</sub>	4.17	0.63	t <sub>self-human</sub> = -3.89**
N <sub>w/human</sub>	3.49	0.54	t <sub>AI-human</sub> = 4.39***
A <sub>self</sub>	5.00	—	t <sub>self-AI</sub> = 14.27***
A <sub>w/AI</sub>	3.54	0.67	t <sub>self-human</sub> = 7.48***
A <sub>w/human</sub>	4.48	0.45	t <sub>AI-human</sub> = -6.36***
E <sub>self</sub>	4.00	—	t <sub>self-AI</sub> = 0.26
E <sub>w/AI</sub>	3.98	0.56	t <sub>self-human</sub> = -3.07**
E <sub>w/human</sub>	4.20	0.43	t <sub>AI-human</sub> = -1.81#
C <sub>self</sub>	4.00	—	t <sub>self-AI</sub> = 4.12***
C <sub>w/AI</sub>	3.63	0.59	t <sub>self-human</sub> = -3.47**
C <sub>w/human</sub>	4.33	0.62	t <sub>AI-human</sub> = -4.32***
SD <sub>self</sub>	5.12	—	t <sub>self-AI</sub> = 15.84***
SD <sub>w/AI</sub>	3.09	0.84	t <sub>self-human</sub> = 7.99***
SD <sub>w/human</sub>	4.28	0.69	t <sub>AI-human</sub> = -5.60***
Control <sub>self</sub>	4.30	—	t <sub>self-AI</sub> = -0.99
Control <sub>w/AI</sub>	4.40	0.68	t <sub>self-human</sub> = 5.08***
Control <sub>w/human</sub>	3.86	0.57	t <sub>AI-human</sub> = 5.28***

\*p &lt; .05; \*\*p &lt; .01; \*\*\*p &lt; .001.

## References

- Ackerman, E. (2016, August 2). *SAM brings much-needed robotic assistance to senior living facilities*. ISPR. Retrieved from <http://ispr.info/2016/08/05/robot-concierge-sam-brings-presence-based-assistance-to-senior-living-facilities/>.
- Adler, A. (1930). Individual psychology. In C. Murchison (Ed.), *Psychologies of 1930*. Worcester, MA: Clark University Press.
- Amalberti, R., Carbonell, N., & Falzon, P. (1993). User representations of computer systems in human-computer speech interaction. *International Journal of Man-Machine Studies*, 38, 547–566.
- Anderson, M. R. (2016, October 17). *Robot babies from Japan raise all sorts of questions about how parents bond with AI*. ISPR. Retrieved from <http://ispr.info/2016/12/22/robot-babies-from-japan-raise-allsorts-of-questions-about-how-parents-bond-with-ai/>.
- Bingblog. (2014). *Meet little ice, Cortana's little sister*. Retrieved from <https://blogs.bing.com/search/2014/09/05/meet-Little.Ice-cortanas-little-sister/>.
- Bingblog. (2016). *Bing helps developers connect more naturally with people*. Retrieved from <https://blogs.bing.com/search-quality-insights/October/bing-helps-developers-connect-more-naturally>.
- Burgoon, J. K., & Jones, S. B. (1976). Toward a theory of personal space expectations and their violations. *Human Communication Research*, 2, 131–146.
- Clark, H. (1996). *Using language*. Cambridge: Cambridge University Press.
- Cooper, C. A., Golden, L., & Socha, A. (2013). The big five personality factors and mass politics. *Journal of Applied Social Psychology*, 43(1), 68–82.
- Duffy, B. R., & Zawieska, K. (September, 2012). Suspension of disbelief in social robotics. In *IEEE Ro-Man: The 21st IEEE International Symposium on Robot and Human Interactive Communication* (Paris, France).
- Edwards, C., Edwards, A., Spence, P. R., & Shelton, A. K. (2014). Is that a bot running the social media feed? Testing the differences in perceptions of communication quality for a human agent and a bot agent on twitter. *Computers in Human Behavior*, 33, 372–376.
- Fischer, K., Foth, K., Rohlfing, K., & Wrede, B. (2011). Mindful tutors: Linguistic choice and action demonstration in speech to infants and a simulated robot. *Interaction Studies*, 12(1), 134–161.
- Franceschi-Bicchieri, L. (2016, September 6). *Porn chatbot tricks Argentinians into thinking they're chatting with President*. ISPR. Retrieved from <http://ispr.info/2016/09/07/porn-chatbot-tricks-argentinians-into-thinking-theyre-chatting-with-president/>.
- Gunkel, D. (2012). Communication and artificial intelligence: Opportunities and challenges for the 21st century. *Communication+1* (Vol. 1)(1).
- Guzman, A. (2016). The messages of mute machines: Human-machine communication with industrial technologies. *Communication +1*, 5(4).
- Hall, J. A., Pennington, N., & Lueders, A. (2014). Impression management and formation on facebook: A lens model approach. *New Media & Society*, 16(6), 958–982.
- Hobbs, J. R., & Evans, D. A. (1980). Conversation as planned behavior. *Cognitive Science*, 4, 349–377.
- Hofstede, G. (1984). *Culture's consequences: International differences in work-related values*. Newbury Park, CA: Sage.
- Hollenbaugh, E. E., & Ferris, A. L. (2014). Facebook self-disclosure: Examining the role of traits, social cohesion, and motives. *Computers in Human Behavior*, 30,

- 50–58.
- Kanda, T., Miyashita, T., Osada, T., Haikawa, Y., & Ishiguro, H. (2008). Analysis of humanoid appearances in human-robot interaction. *IEEE Transactions on Robotics*, 24, 725–735.
- Kelion, L. (2013). Bots now 'account for 61% of web traffic'. Retrieved from <http://www.bbc.com/news/technology-25346235>.
- Kelly, G. A. (1955). *The psychology of personal constructs*. New York: Norton.
- McCord, D. M. (2002). M5–50 questionnaire. Retrieved from <http://paws.wcu.edu/mccord/m5-50/>.
- McCrae, R. R., & Costa, P. T. (1985). Updating Norman's "adequacy taxonomy": Intelligence and personality dimensions in natural language and in questionnaires. *Journal of Personality and Social Psychology*, 49(3), 710.
- McDowell, Z. J., & Gunkel, D. J. (2016). Introduction to machine communication. *Communication +1* (Vol. 5)(1).
- Miller, L. C., Berg, J. H., & Archer, R. L. (1983). Openers: Individuals who elicit intimate self-disclosure. *Journal of Personality and Social Psychology*, 44(6), 1234–1244.
- Mischel, W. (2004). Toward an integrative science of the person. *Annual Review of Psychology*, 55, 1–22.
- Mischel, W., & Shoda, Y. (1995). A cognitive affective system theory of personality: Reconceptualizing situations, dispositions, dynamics, and invariance in personality structure. *Psychological Review*, 102, 246–268.
- Nass, C., Fogg, B. J., & Moon, Y. (1996). Can computers be teammates? *International Journal of Human Computer Studies*, 45, 669–678.
- Nass, C., & Lee, K. M. (2001). Does computer-synthesized speech manifest personality? Experimental tests of recognition, similarity attraction, and consistency attraction. *Journal of Experimental Psychology: Applied*, 7, 171–181.
- Nass, C., & Moon, Y. (2000). Machines and mindlessness: Social responses to computers. *Journal of Social Issues*, 56, 81–103.
- Nass, C., Moon, Y., & Green, N. (1997). Are machines gender neutral? Gender-stereotypic responses to computers with voices. *Journal of Applied Social Psychology*, 27, 864–876.
- Nass, C., Steuer, J., & Tauber, E. R. (1994). Computers are social actors. *Human Factors in Computing Systems*, 94, 72–78.
- Reeves, B., & Nass, C. (1996). *The media equation: How people treat computers, television, and new media like real people and places*. CA: CSLI Publications.
- Shechtman, N., & Horowitz, L. M. (2003, April). Media inequality in conversation: How people behave differently when interacting with computers and people. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 281–288). ACM.
- Shoda, Y., & Mischel, W. (1998). *Personality as a stable cognitive-affective activation network: Characteristic patterns of behavior variation emerge from a stable personality structure*.
- Shulman, S., Laursen, B., Kalman, Z., & Karpovsky, S. (1997). Adolescent intimacy revisited. *Journal of Youth and Adolescent*, 26(5), 597–617.
- Spence, P. R., Westerman, D., Edwards, C., & Edwards, A. (2014). Welcoming our robot overloads: Initial expectations about interaction with a robot. *Communication Research Reports*, 31, 272–280.
- Steyer, R., Schmitt, M., & Eid, M. (1999). Latent state-trait theory and research in personality and individual differences. *European Journal of Personality*, 13(5), 389–408.
- Stuart, S. (2016 October 2). *How social responses to robots are influenced by cultural background*. ISPR. Retrieved from <http://ispr.info/2016/10/12/how-social-responses-to-robots-are-influenced-by-cultural-background/>.
- Sundar, S. S., Jia, H., Waddell, T. F., & Huang, Y. (2015). Toward a theory of interactive media effects (TIME): Four models for explaining how interface features affect user psychology. In S. S. Sundar (Ed.), *The handbook of the psychology of communication technology*. Malden, MA: Wiley.
- Takayama, L., & Pantofaru, C. (2009). Influences on proxemics behaviors in human-robot interaction. In *Proceedings of intelligent robotic systems: IROS 2009, St. Louis, MO*.
- Walther, J. B., Van Der Heide, B., Ramirez, A., Burgoon, J. K., & Peña, J. (2015). Interpersonal and hyperpersonal dimensions of computer-mediated communication. In S. S. Sundar (Ed.), *The handbook of the psychology of communication technology*. Malden, MA: Wiley.
- Wiener, N. (1988). *The human use of human beings: Cybernetics and society*. Boston: Da Capo Press.



# Ethics, Human Rights, the Intelligent Robot, and its Subsystem for Moral Beliefs

Erik Sandewall<sup>1</sup>

Accepted: 1 March 2019 / Published online: 11 March 2019  
© The Author(s) 2019

## Abstract

The Universal Declaration of Human Rights specifies a number of properties that characterize human beings, such as ‘dignity’, ‘conscience’, and several others. In this article we focus on these properties and on how they have been defined in the history of philosophy. We show how they can be interpreted in terms of a prototypical architecture for an intelligent robot, and how the robot can be provided with several aspects of ethical capability in this way. The key idea is to provide the robot with a Moral Belief System that cooperates with, and moderates the robot’s capability of planning and action.

**Keywords** Robot ethics · Moral belief state · Giovanni Pico della Mirandola · Immanuel Kant · Universal Declaration of Human Rights

## 1 Introduction

The present investigation started with the question whether the Universal Declaration of Human Rights (UDHR) [1] could be applied to intelligent robots. If this is feasible, then it may contribute both to the discussion of how governments should relate to such robots, and to the definition of behavior rules for them. The UDHR states that the freedoms that it claims for all human beings must not be used by them as a license to violate the rights of others. This clearly translates into restrictions on the appropriate behavior of agents (humans or robots) that may be covered by the UDHR. In addition, since the UDHR makes a number of statements about properties that humans have intrinsically, we felt that it would be interesting to relate those statements to design principles for intelligent robots.

This curiosity-driven approach turned out to be quite rewarding, in particular when applied to the concept of intrinsic properties of humans. The UDHR uses a few such key concepts, in particular ‘dignity’, ‘reason’, and ‘conscience’, but also several others. These are classical concepts in philosophy which have been extensively studied there. The task of understanding that whole literature, with the various opinions on its topic, was not within reach. We decided therefore

to identify the original authors whose work has set the direction for many of the subsequent contributions, and to study their definitions for the key concepts in some detail.

This choice led us to consider two authors in particular, namely Giovanni Pico della Mirandola who wrote the foundational article ‘Oration on the Dignity of Man’ [2] in 1486, and Immanuel Kant whose works include ‘Critique of Pure Reason’ [3] and ‘Critique of Practical Reason’ [4]. Both of these are very relevant for understanding the key concepts in the UDHR.

But it turned out that besides shedding light on the intended meanings of the statements in the UDHR, the concepts that were proposed by these authors could as well be related directly to the design principles for intelligent robots. At the same time, we also consulted the documents of the United Nations committee that authored the UDHR [5], in order to ascertain that their assumptions and views were consistent with the ones that we obtained from the writings of Pico and Kant.

With all due respect for the authors of the UDHR, however, for the present article we feel that it is appropriate to base it on the definitions of Pico and Kant, as well as other original authors. Our article is therefore organized as follows. We shall first define our assumptions about the design of intelligent robots, in terms of an idealized architecture that specifies what components must be present in it, a few optional components, and the relationships between the components. Next, we identify what we consider to be the essential concepts

✉ Erik Sandewall  
erik.sandewall@liu.se

<sup>1</sup> Linköping University, Linköping, Sweden

in the UDHR for our purpose, and explain why some other concepts have not been included. After this, we address one concept at a time and discuss its interpretation by our classical authors, or by others. In these considerations we are also led to introduce a few additional concepts (besides those mentioned in the UDHR) that are important for seeing the full picture.

In the second half of the article, we propose an additional component for an intelligent robot, which we call its *Moral Belief System*. This component consists of a collection of moral beliefs (the Moral Belief State), together with software that is capable of applying those beliefs in the operation of the intelligent robot as a whole. The purpose of the Moral Belief System shall be to modify the robot's behavior so that it conforms as well as possible to 'moral rules' in a sense that will be further discussed below. The acronym MBS will be used for both the system in question, and for the belief state that it contains.

In summary, there were two reasons for the idea to identify (proposed) intrinsic properties of humans, and to see whether and how they can be applied to intelligent robots as well. The first reason on our part was mere curiosity, but as the work proceeded, we began to see interesting implications for the design of an ethical capability in those robots. One may speculate whether the resulting, operational definition of ethical competence may also be of interest for scholars in the field of ethics, but we must leave it to them to have an opinion on this.

## 2 Relation to Existing Work on Machine Ethics

The present article combines a grounding in classical philosophical concepts of ethics with an explicit model of the intelligent agent. This represents a step forward relative to earlier publications. The first generation of thoughts about the ethics of robots originated with Isaac Asimov's 'Three Laws of Robotics' [6], and was dominated by reasoning about various proposed laws, in particular, their necessity or plausibility, and their consequences.

Contemporary work on robot ethics started with Michael Anderson's and Susan Leigh Anderson's seminal article from 2007: 'Machine Ethics: Creating an Ethical Intelligent Agent' [7]. In this article they discuss the importance of machine ethics, the need for machines that represent ethical principles explicitly, and the challenges facing those working on machine ethics. They also give a simple example of how a machine may abstract an operational ethical principle from examples of correct ethical judgments. Subsequent work has mostly continued the discussion of these topics, or demonstrated simple examples of ethical reasoning in a machine.

With respect to theory, deontic logic is an obvious candidate for a theoretical foundation for machine ethics. However, it can only be used if it is shown how to integrate it into an architecture for the intelligent robot as a whole. Our approach in the present article is to begin with that problem.

In the discussion about machine ethics, a few authors such as Patrick Chisan Hew have expressed doubts about its usefulness, arguing that "Such systems are a substantial departure from current technologies and theory, and are a low prospect" [8]. We feel on the contrary that substantial departures from current technologies have occurred repeatedly in information technology, and we see no reason why one could not happen in the present case. We propose also that the design considerations in the present article may be a step in this direction.

One example of a 'substantial departure', of a very different kind than the one suggested in this article, was proposed in a book edited by Stephen Palmquist, "Cultivating Personhood: Kant and Asian Philosophy" [9]. This book contains an extended argument about whether and how 'reason' and 'free will' can arise from the biological processes of conception and fetus development. This is an example of the broad range of contributions to the topic of machine ethics.

## 3 Intelligent Robots

For the purposes of the present article, an intelligent robot is a mechanical device that at least is equipped with sensors, actuators, a sensory-motoric system and a planning and action system (PAS). The sensory-motoric system contains driver software for the sensors and actuators, and a perception capability whereby a concise description of the robot's current environment is obtained from the sensor inputs, and is also updated continuously. This description will be referred to as the robot's *world model*. In simple cases, the world model will just represent a few objects in the robot's field of observation, with some features of these objects and some relations between them. In more advanced systems, the world model may also represent remote objects, complex objects, present and past events and actions, and other constructs as studied in the field of Representation of Knowledge.

The Planning and Action System is responsible for the robot's choice of actions, the execution of those actions, and various kinds of reasoning about actions. Several parts of this PAS will be relevant for our discussion of robot ethics, and we shall therefore state our assumptions about it with some detail. Our assumptions are consistent with system architectures that are used in practical systems; see e.g. [10].

The PAS shall include a set of designators for *elementary actions* and likewise a set of designators for *composite actions*. Each elementary action designator is associated with a computational mechanism for performing the action. This

mechanism may be defined as a procedure in a programming language, but a number of other design paradigms are also possible, for example in terms of finite automata.

Each composite action designator may be associated with one or more plans that can be used for performing the action. In simple cases, a plan is a sequence of (lower level) action designators. More complex plans can be formed using conditional and repetition operators, as usual.

Furthermore, each action designator shall be associated with information about how the execution of the action is expected to affect the state of the robot's environment, in terms of the categories that are used by the sensory capability. Based on this, the PAS shall have a *prediction capability* whereby it can compute an expectation of the result state after performing an elementary or composite action.

Finally, the PAS shall contain an *action suggesting mechanism* which is able to generate or select one or more action designators based on the current world model as generated by the perception capability. It may be implemented, for example, as a set of situation-action rules.

These are the minimal requirements, and additional facilities may be considered. They include a *goal suggesting mechanism*, as well as a *planning mechanism* for constructing a composite action that is expected to lead to a situation where the robot's environment satisfies certain conditions that are referred to as 'the goal'. An additional and useful facility will be an *action evaluation mechanism* that can be applied to the outputs of the action suggesting mechanism, in order to assess the cost, the likelihood of success, and other characteristics of a suggested action.

## 4 Reasoning and Intelligence

With few exceptions, human beings have a prediction capability and an action suggesting mechanism, similar to those that were described above for intelligent robots. Several alternatives come to mind with respect to how these capabilities work. One possibility is that they work by logical reasoning which is performed in a step-by-step fashion, and where the conclusions in each step can be communicated to others and discussed with them.

Another possibility is that they "just happen" in the sense that the person in question is not able to explain how she or he arrived at the prediction or the suggestion at hand. This is essentially the distinction between 'conscious' and 'subconscious' cognitive processes. We shall refer to these alternatives as "reasoning" and "intuition", respectively. They are not mutually exclusive, since it seems that people can use both depending on the situation at hand, and also since a result that was arrived at by "intuition" can later be explained and motivated as if it had been obtained by reasoning.

These human characteristics have immediate counterparts in the design of robots, where "reasoning" can be implemented using logic-based techniques, and where "intuition" can be implemented using neural networks, for example.

With these definitions and premises, we can proceed to the question of whether, and under what conditions may it be possible to apply the Universal Declaration of Human Rights (UDHR) to intelligent robots.

## 5 Key Concepts in the UDHR

Four concepts in the UDHR are of central importance if it shall be applied to intelligent robots, namely 'freedom' (or 'free'), 'dignity', 'reason' and 'conscience'. They appear as follows in its Article 1 of the UDHR:

All human beings are born free and equal in dignity and rights. They are endowed with reason and conscience and should act towards one another in a spirit of brotherhood.

Some of these words also appear as follows in its Preamble:

Whereas recognition of the inherent dignity and of the equal and inalienable rights of all members of the human family is the foundation of freedom, justice and peace in the world...

The interpretation of the phrase 'in a spirit of brotherhood' will be discussed in the section about the concept of 'conscience', later on in this article. Therefore, at this point we only need to consider how the terms 'free', 'dignity', 'reason' and 'conscience' can be applied to intelligent robots. These concepts are treated differently in the other 29 articles. The concept of 'conscience' appears only once, in Article 18 which begins:

Everyone has the right to freedom of thought, conscience and religion...

The concept of 'dignity' appears only in Article 23:

Everyone who works has the right to just and favourable remuneration ensuring for himself and his family an existence worthy of human dignity...

and the concept of 'reason' does not appear anywhere else besides in Article 1. By contrast, the concept of 'freedom' appears repeatedly, in particular since several articles define what specific 'freedoms' shall be enjoyed by everyone. In fact, the words 'free' or 'freedom' appear 21 times in these 29 articles.

Some other frequently occurring concepts are 'rights' and 'entitled to'. For example, Article 28 says:

Everyone is entitled to a social and international order in which the rights and freedoms set forth in this Declaration can be fully realized.

I shall not address these concepts here, since they are mostly used for ‘rights’ that must be guaranteed by national governments, which means that they must be considered in the broader context of how governments relate to intelligent robots. This would take us outside the scope of the present article.

It remains, therefore, to consider the concepts of ‘freedom’, ‘conscience’, ‘reason’ and ‘dignity’ as used in the UDHR, and for each of them to discuss its requirements on the design of an intelligent robot. With respect to conscience, for example, we need to discuss what may be needed for the robot to have a conscience. Ideally we would also like to understand how the robot may recognize and respect the conscience of others, but this is outside the scope of the present article. I shall address these four concepts one at a time, together with a few others that will come up along the way.

## 6 The Concept of Freedom

The UDHR specifies a number of freedoms for everyone, but it is clearly not an exhaustive list of all the freedoms that people have. Therefore, at first sight, it would seem that the ‘freedom’ aspects of the UDHR should not constrain the permissible behaviors of the intelligent robot at all. However, the Declaration also specifies certain restrictions on the freedoms, namely, in the clauses of Article 29:

- (1) Everyone has duties to the community in which alone the free and full development of his personality is possible.
- (2) In the exercise of his rights and freedoms, everyone shall be subject only to such limitations as are determined by law solely for the purpose of securing due recognition and respect for the rights and freedoms of others and of meeting the just requirements of morality, public order and the general welfare in a democratic society.
- (3) These rights and freedoms may in no case be exercised contrary to the purposes and principles of the United Nations.

The first clause gives a lot of room for interpretation: what is included in the “duties to the community”, and how can a decision or a consensus be reached on such issues? The second clause states that freedoms can only be restricted by law, but in actual practice one would prefer for intelligent robots to recognize informal requirements of morality and public order, and not merely the legal prohibitions in force. In this case the room for interpretation increases even further.

Therefore, in spite of the emphasis on ‘freedom’ in the UDHR, it also allows for restrictions on those freedoms, and the extent of these restrictions is a matter for the society where the UDHR is applied. This is also related to the question of ‘conscience’ which will be addressed later on in this article.

Somewhat surprisingly, therefore, the implications of the UDHR for the design of intelligent robots impose important *restrictions* on the robot’s freedom of action.

## 7 The Concept of Dignity

The ‘dignity’ of all human beings is emphasized very strongly in the Preamble and in Article 1, but UDHR does not even begin to define what it means. It is a well-established term in moral philosophy, however, and a brief review of its origins is therefore in place here. Its Latin form, ‘dignitatis’, is derived from the adjective ‘dignum’; a person is ‘dignum’ of doing something if he or she deserves to do it and is competent of doing it. This noun was promoted by Giovanni Pico della Mirandola in [2] where he proposes that humans have a particular ‘dignity’ because they have free will, whereas both physical objects and animals can only react to whatever forces operate on them. Because of their capacity for free will, humans are able to change themselves, thereby ascending in a chain that leads from the physical world at the bottom, to the divine world at the top.

The concept of dignity was also adopted by Immanuel Kant in order to make a distinction between things that can be discussed merely in terms of what value they have for someone, and those that can not or must not be considered only in those terms. The requirement on the latter is that they shall have a moral dimension, which meant for Kant that they are ‘ends in themselves’. He wrote: “Morality, and humanity insofar as it is capable of morality, is that which alone has dignity” [4,11]. Like Pico, he also considered free will to be an essential aspect of the dignity of the human person.

The combination of free will and morality is therefore of interest for our topic of study. Two later sections will address these two topics in succession. The next section after them will introduce the concept of a Moral Belief System, which we propose as a facility in a robot whereby it can use its morality to voluntarily constrain its own free will. In this way we address the question of how morality and free will can be combined operationally, which is such a crucial issue for our two classical authors.

## 8 The Concepts of Reason and of Being Endowed with Reason

The UDHR states that all human beings are “endowed with reason”. This phrase is not in common use, but probably most readers have an informal understanding of what it may

mean. A Google search for this phrase has only returned references to the UDHR, to a few texts of religious origin, and to works of Kant or about him. In particular, in [11], Kant uses this phrase synonymously with “being a rational being”, which agrees well with common-sense interpretations of these phrases.

At the same time, the concept of ‘reason’ itself is of paramount importance in Kant’s philosophy, as one can see already by its appearance in the titles of several of his major works [11,12]. Kant makes an interesting distinction between ‘logical’ and ‘pure’ uses of reason. The former involves the use of logic for drawing conclusions, which Kant considers as a ‘subordinate’ faculty. In the pure use, on the other hand, “reason itself contains the origin of certain concepts and principles” [11]. Logical reason can therefore be implemented by reasoning software in a computer, whereas pure reason is bound to be a design principle for the software and for the representation of knowledge in the robot.

However, Kant’s works are written in German and use the term ‘Vernunft’ which is then translated into ‘reason’ in English. This translation is treacherous since ‘Vernunft’ means ‘reason’ in the sense of being sensible whereas it does not have any significant connection to ‘reasoning’.

For the present purpose, we shall combine Kant’s two varieties of ‘reason’ and say that a computational agent “has reason” if its software and knowledge representation are organized in a rational way and if they are consistent with pure reason in the sense of Kant. In practice, it is likely that logical reasoning will be extensively used in its Planning and Action System, but this will not be essential for it ‘being endowed with reason’. The first part of this condition means for example that when the PAS initiates the execution of an action, it only chooses an action that has been obtained from the action suggesting mechanism. It also means that the planning mechanism will only produce plans that are expected to achieve the given goal.

## 9 The Concept of Free Will

The question whether robots can be said to have free will has been the topic of much debate in recent years. However, already in the year 2000, John McCarthy published an article with the title “Free Will—Even for Robots” [12] that effectively settled the issue. McCarthy adopted the philosophical stance of compatibilism [13], a philosophical line of thought that goes back to the classical stoics. In this view, free will and determinism are mutually compatible, so it is possible to embrace both without being logically inconsistent. An instance of ‘free will’ may be seen as one in which the agent was able to choose

between alternative goals or actions according to its own internal, cognitive processes, even though these processes may appear entirely deterministic for an outside observer. Kant expressed this view by saying that a rational will cannot act except “under the idea” of its own freedom [11].

In his article, McCarthy also proposed to view ‘free will’ as a graded concept, so that the ‘will’ of an agent may be considered as more or less free, depending on whether and to what extent the agent has internalized external constraints on its behavior. For example, if a parent has forbidden a child to perform a particular action, then the child will be said to have less free will if it has adopted the restriction as its own, and more free will if the child is mentally prepared to perform the action in spite of the parent’s instructions. The latter case applies even if the child does not actually perform the action, for example for some other reason that arises independently of the parent.

The concept of dignity according to Pico della Mirandola and according to Kant is strongly tied both to the existence of free will, and to the presence of morality that influences the choices of this will. This is a strong philosophical reason for paying attention to the combination of those two concepts when providing intelligent robots with a sense of ethics. It follows that such an enterprise can only make sense under the compatibilistic view of free will, since computers must be viewed as deterministic devices. (The introduction of a randomization device would not change this matter). The compatibilistic view is therefore the natural choice when discussing the dignity of robots.

## 10 The Concept of Morality

The article about ‘Definition of Morality’ in the Stanford Encyclopedia of Philosophy [14] distinguishes between a descriptive use and a normative use of this word. The normative case is explained as *a code of conduct that, given specified conditions, would be put forward by all rational persons*. This is well in line with the writings of classical authors, such as Kant.

The descriptive use of this term refers to *certain codes of conduct put forward by a society or a group (such as a religion)*. It seems likely that this meaning is intended in Article 29 of the UDHR when it states that freedoms shall be restricted by *the just requirements of morality, public order and the general welfare in a democratic society*. With respect to intelligent robots, therefore, they should implement normative morality in order that they can be considered to have dignity, and they should also have knowledge of morality in the descriptive sense and for the environment that they are in, so that in their actions, they will not violate Article 29.

## 11 The Moral Belief System

We shall now describe an additional, proposed subsystem of the intelligent robot, called the ‘Moral Belief System’ (MBS) containing its verdicts about which actions and situations are ‘right’ and which are ‘wrong’. The primary purpose of the MBS is to allow the robot to decide for itself what are the moral constraints on its own choice of actions, in line with the compatibilistic view that was described above.

In the context of the UDHR, the MBS is important for interpreting Article 29 so that the robot will not perform actions that it should not. The MBS is also needed, in line with Pico’s argument, so that the robot shall be competent to exercise its own free will and so that it shall deserve doing so. And finally, the Moral Belief System should be seen as the carrier of the morality that Kant required as a condition for dignity.

I shall first outline the technical characteristics of the Moral Belief System (MBS), and later on discuss whether and to what extent it can provide the functionality of a conscience in the sense of the UDHR. The MBS is designed to cooperate with the robot’s Planning and Action System (PAS) that was described above, allowing it to recognize when a proposed action may produce ‘bad’ effects, or ‘good’ effects. To this end it must minimally contain value statements that label particular conditions in the robot’s world as being ‘bad’ or ‘good’, maybe with further qualifications or a graded scale. The same range of values may also be assigned to actions themselves. What is now said about actions applies also to goals and plans. Additional features of the MBS will be introduced below.

If an intelligent robot is designed so that in each situation, it addresses the output of the action suggesting mechanism and selects one of the suggested actions, then the MBS as now described may be used as a filter on the suggested actions, allowing only some of them.

On the other hand, in a robot that uses a goal suggesting mechanism, and planning for finding a plan that achieves the selected goal, the combination of the PAS and the MBS will work as follows. When a goal has been selected, the robot shall first verify that this goal is compatible with the Moral Belief State. If it is, the robot must consider possible plans for achieving this effect, and it will use the planning mechanism in order to obtain one or a few proposals for possible plans. According to the normal operation of a PAS, it will consider these proposals from the point of feasibility, cost, and other factors that may be relevant. With an MBS, it will also consider the possible additional effects (‘side-effects’) of each action in a plan, and relate them to the statements in the Moral Belief State. If it is determined that a plan can have side-effects that are unacceptable according to the MBS then that plan must be disqualified or amended.

## 12 Autonomous Modification of the Moral Belief System

The moral stance of a person or a group is not fixed; it can change over time when new facts become known and when circumstances change. For example, if a group has traditionally maintained that homosexuality is an evil thing, and later it realizes that this view has distressing consequences for the people involved, the members of the group may reconsider the reasons for their traditional view, and then change “their MBS” in this particular respect.

This example shows that when a robot is equipped with a Moral Belief System, one may consider an advanced facility where the robot is able to observe actions that are performed by other agents in its environment, and to assess the consequences of those actions according to the value statements in its MBS. In this way the robot will obtain a larger corpus of events as a basis for its deliberations. Some of these events may involve violations of values in the observer’s MBS and cause it to revise its value structure.

The robot’s capability for reviewing and revising its Moral Belief State will be further improved if it is engaged in a continuous dialog with other agents (i.e., other robots, or persons) concerning the events that they observe together. Just like people tend to exchange views about the goodness or badness of specific events, and the arguments for their respective positions, a robot that engages in a similar dialog may be led to revise its Moral Belief State.

Moreover, the example above illustrates the interdependence between actions and value statements: actions have consequences that may be assessed according to the values, but the adoption of values by a person or a group will also have consequences since it affects what actions are taken and what actions are not taken. This means that there is a consistency requirement on the MBS: it should not contain value statements whose presence there has consequences that are considered as bad according to that same MBS.

Accordingly, it is not sufficient to check the Moral Belief State for consistency a single time, since new facts arrive (by observations, or by dialogue with other agents), and they may introduce inconsistencies in a previously consistent MBS. Moreover, the MBS shall not be analyzed merely as a collection of value statements; it must be seen in connection with the facilities of the entire Moral Belief System, i.e. facilities for the assessment of observed facts, for the detection of inconsistencies in the MBS, and for changing it so as to repair such inconsistencies. The software ‘engine’ that is needed for this purpose will resemble a reason-maintenance system [15] which is a classical device in A.I.

One may ask whether it would not be better to design the MBS in such a way that it does not contain any inconsistencies, and so that inconsistencies can not possibly occur as the result of new information, or as the result of autonomous

changes in the world model or the MBS. This will be a very difficult task, however, and the alternative is to design it using current techniques for reasoning in the presence of inconsistencies [16]. This possibility is also supported by the observation that humans seem to manage very well in spite of inconsistencies both big and small.

### 13 General Rules for the Moral Belief State

Besides the case-driven revision of the MBS, there are also some well-known moral principles that can be seen as restrictions that must be satisfied by the MBS as a whole. The ‘golden rule’ of Jesus Nazareus is an example of such a global restriction:

So whatever you wish that others would do to you, do also to them, for this is the Law and the Prophets (Matthew 7:12).

Kant’s principle of universalizability can also be seen as such a restriction, and as a generalization of the golden rule. One of its formulations is as follows [11]:

Act only according to that maxim whereby you can at the same time will that it should become a universal law.

In our terms, a ‘contradiction’ is a situation where the application of the propositions and mechanisms in the robot’s MBS can lead to results that are ‘bad’ according to that same MBS. However, this formulation of the principle is not very practical, since it would require considerable deliberation each time the robot considers performing a particular action. Therefore it would make more sense to use it as a constraint on the combination of the Planning and Action System and the Moral Belief System in each particular robot, as follows: the MBS must only contain value statements that could be adopted by all humans or robots (or by almost all of them) without obtaining any significant ‘bad’ results. Moreover, the same must apply for the situation-action rules in the PAS action suggesting mechanism.

### 14 Autonomous Modification of the World Model

The world model was mentioned initially as one of the essential components of the intelligent robot. Except in very simple cases, the world model may represent physical objects that are of interest for the robot, properties of objects and relationships between them, current actions and events, past events and foreseen future events, and so forth. This world model will be continuously updated according to the robot’s perceptions, and by communication with other agents.

Like in humans, the world model is interdependent with the low-level perception capability, so that perception updates the world model and the world model affects the perception. However, several other facilities in the intelligent robot will also need to use the world model, including in particular the Moral Belief System. For a simple example, if the MBS shall express that a particular action is required, appropriate, or excluded in a particular type of situation, then both the action and the situation description should be expressed in the terms used by the world model. Also, the world model must be used for reasoning about such actions and situations. The same applies for the action suggestion mechanism and the action evaluation mechanism.

Because of these interdependencies, one should also keep in mind the possibility that the Moral Belief System may influence the world model. This may happen, in particular, as the result of a cognitive dissonance, i.e. a mismatch between different desires or inclinations in the robot. When the robot’s cognitive system attempts to remove the reasons for a cognitive dissonance, one possible diagnosis on its part may be that the dissonance is the result of a fault in the world model, which may lead it to reconsider some aspect of that model. This kind of “new understanding” does occur sometimes in humans. Although it is not easily replicated in artificial systems, one should at least attempt to design those systems in such a way that this kind of rethinking is not excluded from the outset.

### 15 The Concept of Conscience

The UDHR states that all human beings are endowed with conscience. It therefore makes sense to ask whether a similar facility would also be needed and useful in an intelligent robot, and whether it could contribute positively to the robot’s behavior. Moreover, as we shall see, the concept of conscience has some interesting implications for the design of the intelligent robot.

The Stanford Encyclopedia of Philosophy describes one meaning of this word as follows:

When we talk about conscience, we often refer to reflection about ourselves as moral persons and about our moral conduct. Through conscience we examine ourselves, as if we were our own inner judge. The image of an individual split into two persons, one who acts and the other who observes the former’s conduct, reflects the original conception of ‘conscience’ in the Greek world, at least from the fifth century BCE.

Let us consider this definition first, and then return to other definitions. The one at hand allows for two operational interpretations, depending on whether the “judgement” of an action occurs before or after it has been performed. In

the first case, the conscience operates as an enabler or a disabler of actions, so that a person's conscience may rule that he or she shall absolutely not perform certain actions, or that in certain situations he or she shall perform certain actions.

This aspect of the conscience is clearly accommodated by the combination of the PAS and the MBS that was described above. It imposes a requirement on the system design, namely, that the morally required actions are included in the output of the action suggesting mechanism.

A problem arises if a person's conscience requires him or her to perform actions that are permitted by the general formulations of the freedoms in the UDHR and supported by the freedom of conscience (Article 18), but which also happen to be restricted according to the local interpretations of Article 29 which was discussed above. This occurs when a person's conscience tells him or her to act contrary to the expectations of the surrounding society. A well-developed MBS should be capable of identifying such clashes, and a well-developed PAS should contain methods for dealing with them.

The other operational interpretation of 'conscience' comes into play when a person experiences having 'a bad conscience' for something that he or she has done, or has chosen not to do. In this case it is a question of retroactively assessing an action of one's own and wishing that one had done otherwise. This can lead to new concrete actions, such as to apologize for example, but it can also lead a robot to reconsider some parts of its MBS, or even some aspects of its decision-making machinery. In other words, the important aspect of conscience in its second function is not the identification of the fault, since this can also be taken care of by the PAS and the MBS. Instead, the capability of changing oneself as a result of the remorse shall be understood as an additional competence that is an integral part of the conscience.

The view of the conscience as an inner judge is a powerful one, but it does not cover all uses of the term. In particular, conscience can also be seen as the basis for obligatory situation-action rules, which say that in a particular kind of situation, the agent must necessarily perform a certain action, regardless of other considerations. This may be related to the concept of identifying the 'Is' and the 'Ought', as advocated by Hallaq [17]. He explains it through an example where the observation that a person Is poor shall be ontologically identified with the requirement that one Ought to help the person. The example is not entirely convincing, but the idea is interesting.

To summarize, a fully developed MBS (consisting of both a software engine and a collection of value statements) should make the robot capable of revising the set of value statements in its MBS, both by plain deliberation and by conscience-driven reconsideration of those statements. This possibility

is very much in line with Pico's view of how the soul can ascend from the physical world and towards the divine world.

Finally, we shall return to the interpretation of the phrase 'in a spirit of brotherhood' which occurs in Article 1 of the UDHR. The website of the UDHR Project at Columbia University defines this phrase as follows, in its page on Article 1 [18]:

To treat one another in a "spirit of brotherhood" means that individuals should, in a figurative or symbolical sense, treat each other in such a way as proper to the relation of a brother.

This explanation leaves a lot to the reader's imagination. For the present purpose I shall assume that a spirit of brotherhood consists, at least, of 'empathy' and 'solidarity'. It may be argued that 'understanding' of the other person's character should also be included as a separate point, but I will leave that aside as being too complicated to deal with. Empathy, then, might be characterized as a capability of the agent's perception system whereby it is able to interpret its observations of others in terms of its model of itself and its own experiences, combined with an action or goal suggesting mechanism that reacts appropriately to these interpreted observations. With this understanding of the phrase, 'empathy' is closely related to the variety of 'conscience' that takes the form of obligatory situation-action rules.

The concept of 'solidarity' may be seen as analogous to 'empathy' but with the difference that it does not rely so much on the observer's model of itself. A person may show solidarity with the needs of his or her brother although they do not experience having the same needs themselves.

These definitions of terms such as 'empathy' and 'solidarity' must be seen as first approximations which can not of course do full justice to these words as they are used in psychology or in politics, and even less when compared to how they are used in literature. However, one must start somewhere, and even these crude operational definitions may be useful for relating terms such as these to the design of intelligent robots whose behavior may have some resemblance to 'empathy' and 'solidarity' in humans.

## 16 Potential Problems with Morality-based Robots

With respect to software systems that have a limited level of "intelligence", it is sometimes said that "limited intelligence is worse than no intelligence at all". Be that as it may, one may worry that an analogous statement may hold some truth: "a system with limited morality and understanding is worse than a system with none of those". After all, the very point with

endowing an intelligent robot with a sense of morality can only be to keep it from doing things that are evil or have evil effects, and to incite it to do things that promote goodness. Furthermore, there is an underlying assumption that the robot shall be enabled to make these judgements autonomously. This can only be worthwhile if the considerations in question are so complicated (or so incompletely known) that they can not be completed in advance. If they can, then one may as well resort to a preprogrammed ethical behavior, without any need for a ‘free will’.

But if the situations that the robot will encounter can not be predicted and circumscribed, and if the robot’s behavior in those situations can not be predicted in advance, then how can we know that the interventions of the robot will have positive effects in all the situations that it encounters? And if we deal with this concern by asserting that the robot’s actions are for the good in the overwhelming majority of cases, although maybe not all of them, then how shall we as a society relate to those situations where the robot’s actions turned out to be quite detrimental? Shall they be considered as accidents that are caused by Nature, or shall someone be considered as responsible—the robot’s designer, the robot’s owner, or the robot itself?

Besides these major problems, there is also an issue even in the case where an intelligent robot is merely used for observation and intelligence-gathering. We have already remarked that there is a possibility that the Moral Belief System can influence the world model in some situations. By way of speculation, one may imagine a robot that has too much morality and autonomy for its common sense; if such a robot is assigned the task of observing public spaces and if it can file a report or send in a task force when it sees that something bad is about to happen—then what? The concept of morality-based surveillance has some obvious problems.

What we have said here should not be used as arguments against the implementation of ethics in robots and other computer based systems. However, it does qualify as arguments for proceeding slowly and carefully when this kind of technology is put into use. Information technology has a deplorable track record of how new technologies and new systems have been put into operation prematurely. We should not let that happen for morality-based robots.

## 17 Alternative Uses of the Considerations Made Here

The present article has focused on the question whether the UDHR’s statements about the nature of humans can be applied to the design of intelligent robots. There is a related and important issue that we have not addressed, namely, how an intelligent robot shall be designed so that it does not violate the human rights of any person in their environment. One

may suggest that this can be done simply by including the articles of the UDHR to the robot’s Moral Belief State, but it seems that this would impose an overly difficult task on the MBS software. It seems more likely, therefore, that the rules of the UDHR shall have to be “compiled manually” into value statements and behavior rules that are suitable for being integrated into the robot’s MBS.

## 18 Alternative Definitions of Dignity and Morality

In this article we have used definitions of dignity and morality that are characteristic of Western culture during the period of modernity. Although our two selected authors embraced Christianity, more or less, their concepts and lines of reasoning did not reflect the religious doctrines at the time.

However, the dignity of the human person is also an important concept in several religions, and this should be mentioned briefly here as a reminder, and so that this aspect is not ignored. In religious frameworks, ‘dignity’ is often seen as a requirement to conform to the religious laws. For example, the Islamic scholar Mohammad-Ali Taskhiri views dignity as a state to which all humans have equal potential, but which can only be actualized by living a life pleasing to the eyes of God [19].

The concept of morality is likewise very important in many religions. For example, the Muslim scholar Wael Halq wrote [17]:

...the discursive world of Islam and its forms of knowledge were pervaded by moral prescriptions and by Sharía-prescribed ethical behavior.

Religious frameworks of this kind can be seen as an alternative or as a complement to secular definitions, such as the largely secular one given by Kant. However, inasmuch as abrahamic religions consider that mankind was created in the image of God:

So God created mankind in his own image, in the image of God he created them; male and female he created them.

(Genesis 1:27), it seems quite unlikely that their followers shall be willing to extend their concepts of dignity and morality to man-made devices, such as intelligent robots.

## 19 Summary and Concluding Remarks

In order to address the role of Ethics and of Human Rights in the design of intelligent robots, we have specified a few basic assumptions about the software architecture of such robots. We have also identified a set of important concepts in ethics,

discussed their definitions, and described how they may be realized in the robot's architecture.

The major assumptions about the architecture of the intelligent robot were that it should contain a sensory-motoric system, a perception capability that produces a world model, and a planning and action system (PAS). We have furthermore proposed the addition of a moral belief system (MBS) that interacts with the PAS. The assumed characteristics of these subsystems have been described in outline.

With respect to the concepts in ethics, we have chosen the Universal Declaration of Human Rights (UDHR) as the starting point for our analysis. It specifies four properties that are stated to be characteristic of human beings, namely ‘freedom’, ‘dignity’, ‘reason’ and ‘conscience’. The concept of freedom is relevant since the UDHR does not merely claim a number of specific freedoms; it also contains clauses that restrict those same freedoms. The other three concepts are important since they characterize those properties of human beings that make us worthy of the freedoms and rights that are defined in the UDHR. Each of these four concepts is therefore relevant for the ethical aspects of intelligent robots.

The discussion of these properties could not be based only on the text of the UDHR and available comments about it. We have therefore included some key ideas of a few foundational authors into our analysis, in particular Giovanni Pico della Mirandola and Immanuel Kant. This also led us to add a few additional concepts into the analysis, in particular ‘free will’ and ‘morality’.

The main result of these considerations is that it provides a conceptual framework for the design of a Moral Belief System that can impart a sense of ethics to an intelligent robot. Furthermore, this conceptual framework may clarify the relationship between the philosophy of ethics and the ethical capabilities of intelligent robots.

The use of the UDHR as the starting point for this work was natural in view of its very broad acceptance, and since it specifies characteristic properties of humans that are relevant for our topic. However, the UDHR also has some weak points. In particular, it talks mostly about ‘freedoms’ and ‘rights’, but only indirectly about ‘obligations’. On the other hand, the analysis of ‘dignity’ as requiring the combination of ‘free will’ and ‘morality’ compensates to some extent for that weak point.

Further analysis of the obligations of intelligent robots would be a natural topic for future work. It is part of a broader topic that has been mentioned above, namely, what will be the appropriate laws and other rules for intelligent robots, and what social expectations should be applied to them.

**Acknowledgements** This article has been much improved thanks to the insightful comments of the three anonymous reviewers.

**Funding** This work was not done as part of any grant.

## Compliance with Ethical Standards

**Conflict of interest** The author declares that he has no conflict of interest.

**Open Access** This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

## References

- United Nations, General Assembly resolution 217 A (1948) Universal Declaration of Human Rights. <http://www.un.org/en/universal-declaration-human-rights/index.html>
- della Mirandola PG (1486/1496) Oration on the dignity of Man (De hominis dignitate)
- Kant I (1781) Critique of pure reason (Kritik der reinen Vernunft). Riga
- Kant I (1788) Critique of practical reason (Kritik der praktischen Vernunft). Riga
- Drafting of the Universal Declaration of Human Rights. A collection of documents. Dag Hammarskjöld Library. <https://research.un.org/en/undhr/draftingcommittee>
- Asimov I (1950) I, Robot. Gnome Press, New York
- Anderson M, Anderson SL (2007) Machine ethics: creating an ethical intelligent agent. *AAAI Mag* 28(4):15
- Hew Patrick Chisan (2014) Artificial moral agents are infeasible with foreseeable technologies. *Ethics Inf Technol* 16(3):197–206. <https://doi.org/10.1007/s10676-014-9345-6>
- Palmquist S (ed) (2010) Cultivating personhood: Kant and Asian philosophy. De Gruyter, Berlin
- Ghallab M, Nau D, Traverso P (2016) Automated planning and acting. Cambridge University Press, Cambridge
- Kant I (1785) Groundwork of the metaphysics of morals. In: Gregor M (ed) Cambridge University Press, pp 53–74. ISBN 9780521626958. OCLC 47008768
- McCarthy J (2000) Free will—even for robots. *J Exp and Theor Artif Intell* 12(3):341–352
- Compatibilism (2002/2015) Stanford encyclopedia of philosophy. <https://plato.stanford.edu/entries/compatibilism/>
- The Definition of Morality (2002/2016) Stanford encyclopedia of philosophy. <https://plato.stanford.edu/entries/morality-definition/>
- Martins J P, Reinfrank M (eds) (1990) Truth maintenance systems. In: Proceedings of the ECAI-90 workshop. Springer lecture notes in computer science
- Johnson-Laird PN, et al. Reasoning about inconsistency—list of relevant publications. Mental models and reasoning. Princeton University. <http://mentalmodels.princeton.edu/portfolio/reasoning-about-inconsistency/>
- Hallaq WB (2012) The impossible state. Islam, politics, and modernity's moral predicament. Columbia University Press, New York
- Danchin P Article 1: Fundamental Human Rights. A page in a website about the Universal Declaration of Human Rights. [http://ccnmtl.columbia.edu/projects/mmt/udhr/article\\_1/meaning.html](http://ccnmtl.columbia.edu/projects/mmt/udhr/article_1/meaning.html)
- Taskhiri M-A (1997) Human rights: a study of the universal and the islamic declarations of human rights. Islamic Culture and Relations Organization, Alhassanain institute, Iran, p 92. <http://alhassanain.org/english/?com=book&id=1153>

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Erik Sandewall** is an emeritus professor at Linköping University, Sweden, and a member of the Swedish Academy of Sciences. His major research interests are in knowledge representation, reasoning about actions, and the use of these topics in the software architecture of

intelligent and autonomous robots. His contributions involve both theoretical work and application projects, in particular concerning the design of intelligent UAV.

# ROBOTS

YOUR GUIDE TO THE WORLD OF ROBOTICS

Home [Robots](#) News Play Learn  ALL ROBOTS SORT ROBOTS ROBOT RANKINGS

Perseverance



Aibo



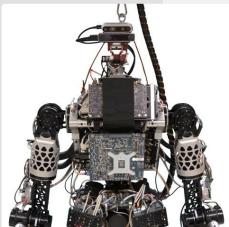
KOOV



TIAGo



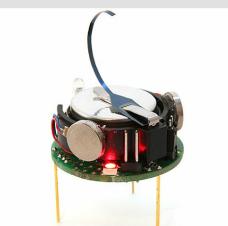
Kamigami



Lola



Charlie



Kilobot



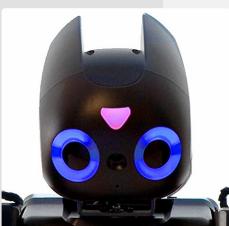
RoboThespian



VGo



AILA



Darwin-OP



Ekso



Hand Arm System



PR2



Robonaut 2



Nao



Rollin' Justin



Pleo



BigDog



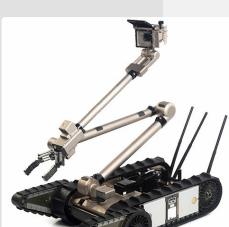
Paro



Roomba



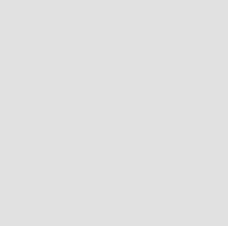
RHex



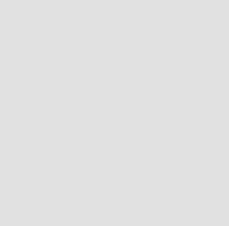
PackBot



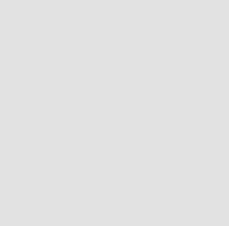
Da Vinci



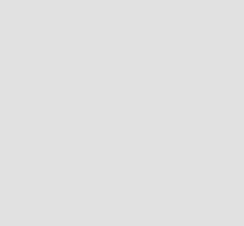
Genghis



Cassie



iCub



Meca500



Baxter

Geminoid DK

Jaco

Drive Unit

Curiosity

Google Self-Driving Car

Albert Hubo

Keepon

Asimo

ACM-R5H

Adaptive Gripper

Aibo (1999)

AirBurr

AlphaDog

Anafi

Anki Drive

ANYmal

Aqua2

Aquanaut

AR-600

ARM

Armar

Atlas

Atlas (2013)

Automatronics

Ava

BallIP

Bandit

Beam

BEAR

BHR-5

Boss

BotVac

Braava

Braava Jet

Bruno

Care-O-bot 4

CB2

Chaos

CHARLI

Chico

Cobalt

Cody

Cog

Colossus

COMAN

Cubelets

Daisy

DASH

Dash and Dot

Diego-san

Digit

Double

DRC-Hubo+

Dreamer

EASE

eBee

ECCE

Elios

Emiew 3

EMYS

Erica

Expliner

Explorer Snake-arm Robot

Fable

Fetch

FLASH

Flipperbot

Flobi

Freight

FROG

Furby

Geminoid F

Geminoid HI

Global Hawk

Guardian GT

Guardian S

HAL

Handle

Harvey

Herb

HEXA

HRP-2

HRP-4

HRP-4C

HRP-5P

Hubo 2

Human Support Robot

Husky

HyQ

i-Limb

iBot 4000

Jibo

K5

Kaspar

KHR-3

Kibo

Kismet

Kiwi

Kobian

Kobra

Kojiro

Kuri

Laikago

LAURON V

LBR iiwa

LD

Lego Boost

Lego Mindstorms EV3

Lego Mindstorms NXT

Lego WeDo 2.0

Leonardo

Lucie

M1

Mabu

Mahru

Mambo FPV

Mavic 2

Mercury

Mini Cheetah

Misty II

Modular Prosthetic Limb

Momaro

Na'vi Shaman

Nano Hummingbird

Nextage

nuTonomy

Octavia	Olivia	OTTO	Partner	Pepper	Perseverance
Petman	Phantom	Photon	Picker Robots	Pioneer 3	Pneuborn
PR1	QB	qb SoftHand Research	Qbo	Qrio	QTrobot
Quattro	Quince	Raven	Raven II	REEM-B	REEM-C
Relay	Replicator+	RoboBee	Roboy	Root	Rosie
RVR	Salamandra robotica II	Sawyer	Segway	Shadow Hand	Simon
Skydio R1	SmartBird	Sophia	Sphero	Spirit & Opportunity	Spot

Stanley Starship Stretch Stuntronics Surena TALOS

Telegarden Telenoid Temi Throwbot Titan TORO

TUlip TURTLE TurtleBot TurtleBot 3 Twendy One UnicornBot

Unimate UR Valkyrie Vector Versatrax Vita

Wabot 2 Wakamaru WAM Waseda Flutist Watson Wave Glider

Waymo YuMi Zeno Zipline

---

FOUNDING SPONSORS

# ROBOTS



About Us   Donate   We ❤ Our Sponsors   Contact Us

[Home](#) | [Sitemap](#) | [Accessibility](#) | [Nondiscrimination Policy](#) | [IEEE Privacy Policy](#)

© Copyright 2022 IEEE – All rights reserved. Use of this website signifies your agreement to the IEEE Terms and Conditions.

A not-for-profit organization, IEEE is the world's largest technical professional organization dedicated to advancing technology for the benefit of humanity.

IEEE websites place cookies on your device to give you the best user experience. By using our websites, you agree to the placement of these cookies. To learn more, read our [Privacy Policy](#).

[Accept & Close](#)