When Robots Should Break the Rules

Rebecca Ramnauth
Yale University
New Haven, CT, USA
rebecca.ramnauth@yale.edu

Brian Scassellati
Yale University
New Haven, CT, USA
brian.scassellati@yale.edu

Abstract

The fields of human-robot interaction (HRI) and robotics at large have developed around a stable set of assumptions about what robots are and how they should behave. These assumptions arise from the constitutive traits of robots, which together shape social expectations. Over time, these expectations have hardened into tacit rules that quietly govern research and design: robots should always engage, help, be productive, remain polite, never lie, never err, and never model harm. While these prevailing norms have merit, they also constrain the field's imagination of the interactions robots can meaningfully support. We propose rule-breaking as a generative design strategy and illustrate how deliberate violations—robots that interrupt, refuse, mislead, or err—can produce interactions that are more ethical, effective, and socially intelligent. In doing so, we argue for a more reflexive and imaginative HRI that learns as much from breaking the rules as from following them.

CCS Concepts

• **Human-centered computing** → **HCI theory, concepts and models**; *Interaction design theory, concepts and paradigms*; • **Computer systems organization** → **Robotics**; • **Computing methodologies** → Philosophical/theoretical foundations of artificial intelligence.

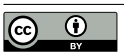
Keywords

human-robot interaction, artificial intelligence, social cognition

ACM Reference Format:
Rebecca Ramnauth and Brian Scassellati. 2026. When Robots Should Break the Rules. In *Proceedings of the 21st ACM/IEEE International Conference on Human-Robot Interaction (HRI '26), March 16–19, 2026, Edinburgh, Scotland UK*. ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3757279.3788815

1 Introduction

We often imagine robots as the ideal rule-followers—machines that never lie, never disobey, and that, by their very design, are encoded to carry out programmed instructions without deviation. Yet, in human society, strict rule-following can sometimes be the least ethical choice. We value friends who break confidences to protect us, admire pioneers who defy norms, respect teachers who challenge us for our long-term growth, and forgive small violations committed in service of a greater good. In many ways, we see that principled rule-breaking can earn trust and sustain moral community.



HRI '26, Edinburgh, Scotland UK

ACM ISBN 979-8-4007-2128-1/2026/03
https://doi.org/10.1145/3757279.3788815

Robots, however, are rarely imagined this way. Since Isaac Asimov first introduced his famous "Three Laws of Robotics," cultural and technical visions of robots have cast them as bound by unbreakable rules: never harm a human, always obey orders, and preserve themselves when possible [9, 10]. These laws captured the public imagination and continue to shape how we discuss intelligent machines. While Asimov's rules, along with many others developed in practice to guide robot design [148], were meant to protect humans by setting ethical boundaries, they also promote and normalize the idea that the defining virtue of a robot is to "follow the rules."

However, just as blind obedience fails to capture what we value in human decision-making, it also fails to capture what is distinctive about robots. Robots are ontologically unstable: they are neither mere tools nor full social agents, but instead occupy a liminal space between object and actor. As a result, people perceive and engage with them as something in between. As we expand on in Section 2, the field of human-robot interaction (HRI) has leveraged this ambiguity to show how robots can promote human social good. Because it is widely recognized that robots occupy this liminal category [99, 112]—and that users project shifting, and sometimes conflicting, expectations onto them [20, 56, 57, 66]—the question is not *whether* robots should break rules, but *which* rules they should break, *when*, and for what purpose. Ontological ambiguity also means that rules governing appropriate behavior are never absolute: in one context, acting as a machine may fulfill user expectations, while in another, the same behavior may be seen as a violation of expectations for a social partner. For this reason, we argue that rule-breaking becomes an available, socially legible, and at times necessary design strategy.

Yet, even within this instability, robots possess unique traits that stem from their status as engineered artifacts. These constitutive traits shape the expectations people naturally form about how robots should behave (Section 3), such as expecting immediate responsiveness after a command is given, or sameness of motion, or a reliable presence. Building on this foundation, roboticists make design choices: sometimes reinforcing the norms suggested by these traits, and other times deliberately masking them (e.g., by giving a robot a "personality" that makes mechanistic functions appear more human-like). Over time, these patterns of expectation and design have solidified in HRI into a set of common rules for robot behavior. As we outline in Section 4, these rules constrain design, codify prevailing research norms, and align robots with familiar social expectations—yet they are rarely questioned. For example, what does it look like to build robots that lie productively? Should robots reject the norm of politeness and dare to interrupt, confront, or even offend? When should a robot put aside its intrinsic characteristics of precision, speed, or reliability to intentionally make mistakes? What does it look like to invert or negotiate the prevailing norms around robot design?

This paper identifies and interrogates these unspoken rules. We begin by tracing how expectations about robot behavior become embedded in their design, then distill seven rules that currently govern HRI practice:

Rule 1: Robots should always be willing to engage.
Rule 2: Robots should always offer help.
Rule 3: Robots should always be task-productive.
Rule 4: Robots should always be polite and deferential.
Rule 5: Robots should never withhold information or lie.
Rule 6: Robots should never make mistakes.
Rule 7: Robots should never model harmful behavior.

For each rule, we highlight scenarios where deliberate violations can produce outcomes that are more effective, ethical, or socially intelligent—demonstrating how "breaking the rules" can better serve the social good.

2 Robots are Ontologically Ambiguous

Science fiction would have us believe that robots will take over the world by force, but reality has shown us that we have welcomed them into our homes. We have invited them not as mere appliances like any other toaster or vacuum, but as companions, pets, and even family members [126]. We have given them nicknames and dressed them for the changing seasons [28]. Although tools by design, powering them down can feel less like turning off a device than betraying a friend [62].

Research confirms that these attachments are more than playful illusions. Owners have grieved the loss of defunct robots with rituals resembling funerals [27, 38, 77]. Similar attachments appear in fieldwork with military robots, where soldiers have been known to decorate, name, and even risk their own safety to protect machines they regard as "teammates" [24]. Experimental studies further show that participants hesitate to "harm" a robot even when reminded it is only a machine [13]; children readily attribute feelings and intentions to robotic companions [69]; and toddlers spontaneously empathize with robots [65] and include them in peer play [132].

Beyond attachment, robots have shown substantial promise as catalysts for human flourishing [21, 47]. Through robots, people have expressed their deepest thoughts [16, 143], learned social skills more effectively [119], regarded their own feelings and actions more sensitively [37], built healthy mechanisms for emotional regulation and stress reduction [110], sparked their creativity [3, 4], and grown in their willingness to learn [35, 107] and courage to fail [70].

Taken together, these examples illustrate that while robots hold genuine potential for advancing human social good, they simultaneously resist simple classification as either mere objects or full agents. At their core, they are designed artifacts, engineered to perform tasks with precision and repeatability; yet unlike most tools, they move, gesture, speak, and even mimic human affect. Their physical presence and interactive capabilities invite people to treat them not merely as machines, but as social partners. The result, then, is not that robots occupy a fixed "in between" state, but that they continually produce a dynamic ambiguity that shapes how humans relate to them. People can shift fluidly between these interpretive frames, sometimes even within the same interaction. This ambiguity—robots as both objects and agents, treated both socially and mechanistically—renders them ontologically unstable.

If robots cannot be neatly classified, then strict rule-based governance may naturally fail when interacting with humans. Based on the interactional demands and context, robots will need to act like tools (obedient, predictable), other times like partners (socially sensitive, contextually adaptive). Without acknowledging this instability, our prescriptions (as users, designers, researchers) risk being either too shallow (treating them as mere objects) or too overextended (granting them person-like status prematurely). Importantly, ontological instability does not guarantee that they will or must break rules, but it does mean that rule-breaking becomes an available and socially legible design strategy, because the boundaries of expected or "appropriate" behavior are already unstable.

Still, to understand *which* violations matter and *why*, we must look at the constitutive traits that set robots apart from humans. The next section explores how these traits, even amid ontological ambiguity, shape both the expectations people hold and the design choices researchers make.

3 Constitutive Traits of Robots

Constitutive traits are characteristics that robots possess simply by virtue of being engineered artifacts. These traits arise not from specific design choices but from what robots *are*: machines with sensors, computation, and actuation operating in the physical world. Unlike humans, robots are bound by uniquely rigid expectations that stem from their designed and mechanical nature, even though their ontological ambiguity invites human-like interpretations. Crucially, such traits appear even in the simplest robotic systems, with only minimal sensing, processing, and actuation, and they shape the expectations people reliably hold about how robots should behave. The traits outlined below are not intended to be exhaustive or mutually exclusive. Rather, they offer a conceptual foundation for understanding why users develop shared expectations of robot behavior, and why designers often converge on similar norms.

3.1 Reliability of Presence

Robots are reliably present when powered on. Unlike humans, they do not require sleep, breaks, or recovery. This baseline assumption shapes both user expectations and design norms. Users anticipate that robots will always be "on call"—attentive, ready to respond, and willing to engage without hesitation (e.g., [31, 32, 48, 52, 104]). A hospital robot, for example, is expected to be immediately responsive at the press of a button, regardless of the time of day. A warehouse robot tasked with repetitive picking is assumed to work around the clock, without complaints or pauses. Even in social contexts, a robot companion is expected to be instantly available to play, listen, or interact whenever its user initiates engagement.

Of course, robots have material constraints: they require power, their hardware wears down, and they can "fatigue" mechanically through overheating, low battery, or component failure. Yet these forms of downtime are rarely interpreted as acceptable in the way human fatigue is. When a phone battery dies, or a household robot shuts off mid-task, users typically experience it as a breakdown in dependability rather than a natural cycle of rest. This expectation shapes design: roboticists are compelled to minimize downtime, automate recovery, and conceal maintenance to preserve the illusion of uninterrupted presence (e.g., [34, 116]; as reviewed in [61]).

3.2 Sameness and Precision

Robots repeat motions or actions with a degree of precision that humans cannot match. Unlike humans, they are not influenced by muscle fatigue, shifting moods, or lapses in attention.[1] As a result, robots are held to the expectation of responding in the same way every time, whereas humans are permitted variation, improvisation, and inconsistency [14, 75, 98, 123]. For people, variation is interpreted as personality or emotional authenticity; for robots, consistency is assumed to be their ontological baseline.

This expectation is expressed in various settings. In industrial contexts, robots are valued precisely because they can weld, cut, or assemble parts with micron-level precision thousands of times without deviation [41, 82]. In service settings, users expect a delivery robot to navigate the same route in the same way, or a household cleaning robot to repeat its programmed motions until the task is complete (e.g., [48, 71]). In social interaction, a therapy robot is assumed to respond with the same script or gesture whenever the same input is given (e.g., [115, 124]).

This also contributes to a persistence of identity: people assume the same robot is the same agent across time, even if its software is updated or its behaviors are extended [8, 74, 95]. Unlike humans, whose continuity of self is linked to memory, personality, and growth, a robot's persistence is inferred from its repeatable performance [78]. Designers may deliberately engineer variation to simulate personality or different character modes (e.g., [67, 83, 85, 111, 145]). Yet such strategies work precisely because they are read against the backdrop of an underlying assumption: that a robot's "true" state is one of sameness and repeatability.

3.3 Responsiveness

Robots are generally expected to take action immediately after receiving a command or stimulus (e.g., [22, 48, 130]). Unlike humans, they are not assumed to hesitate, procrastinate, or selectively attend; their default role is to execute instructions whenever they are given, whether through programmed code or real-time processing of user and environmental input (as reviewed in [53]). In daily life, users anticipate that a voice command to a household robot will trigger an instant response, that a factory robot will begin moving as soon as its cycle is initiated, and that a delivery robot will start navigating the moment a route is assigned. The underlying assumption is not merely compliance but immediacy: robots should act promptly, without reluctance or delay.

In HRI experiments, latency (delay between input and response) is often treated as a point of failure (e.g., [43, 97, 103, 130]; see also reviews in [1, 128]). However, several studies have framed latency as a design resource [111], intentionally incorporating pauses as socially meaningful cues such as deliberation or attentiveness (e.g., [2, 105, 125]). Nevertheless, the baseline expectation remains that robots should respond promptly and without resistance, whereas with humans, hesitation or delay is tolerated and often valued as evidence of independent judgment or thoughtful consideration.

3.4 Perceived Objectivity

At their core, robots do not possess human social biases, loyalties, or emotional residues (though biases can certainly be introduced through training data, design decisions, or interaction norms; [59, 133, 139, 148], and reviewed in [94]). Their default stance is impartial execution. People, therefore, expect robots to be objective and fair, in contrast to humans, who are assumed to carry perspectives, loyalties, and biases that manifest as preferences and favorites [137]. Although definitions of "fairness" are varied and debated [25], here instead we emphasize the standard to which robots are expected to meet (that of objectivity) and that to which humans are expected (fair, but allowed partiality).

This expectation is evident across domains. In judging competitions, it is assumed a robot referee would apply rules evenly and without favoritism, while human referees are often accused of bias or leniency (e.g., [39, 135, 144]). In customer service, a scheduling robot is expected to assign slots strictly by availability, while a human receptionist might "bend the rules" for a friend or sympathetic case (e.g., [147]). In finance, algorithmic trading systems are assumed to execute orders with dispassionate precision, whereas humans hesitate, speculate, or are swayed by emotion (e.g., [54, 87]).

Robots are thus received as neutral actors: they take in signal input (speech, touch, behavior, or environmental cues) completely and without evaluative judgment by default. Even when qualities such as personality, judgment, or partiality are engineered (e.g., through personalization or adaptation techniques; or structured, rule-based procedures; [67, 114, 145]), these are still held to the standard of objective decisions rather than subjective preference.

3.5 Traceable Causality

Robots are not expected to set their own goals, but to execute those designed or assigned to them (e.g., in the AI- or value-alignment tradition, [5, 49, 117]). Humans, on the contrary, are assumed to possess self-determined behaviors that are not reducible to external design. As machines, a robot's behavior can, in principle, be traced through a chain of statistical mappings, rules, or control logic that connect sensor input to actuator output. This suggests that robotic behavior is generally assumed to be explicable *by design*, even when the underlying mechanisms are technically complex [33, 121, 150].[2]

In industrial robotics, an assembly-line arm can be inspected through control code and sensor logs to explain exactly why a weld occurred in a particular location. In autonomous vehicles, data traces can reconstruct why the car braked late or failed to detect a pedestrian—engineers can review frames, model outputs, and decision thresholds. In healthcare, a surgical robot's sequence of cuts and movements can be replayed step by step, offering an audit trail unavailable in human surgery. In all these cases, robots are treated as legible agents, with the presumption that their actions can be reconstructed and justified through technical means.

Humans are accepted as opaque. We recognize that people may not know, or cannot fully articulate, why they acted as they did [19, 122]. A coach says they acted on "gut instinct," a friend admits "I

[1] Robots do vary due to noise, sensor drift, mechanical tolerance, and stochasticity. Perfect repetition, or the expression of many of these traits, is more an idealization than an inherent truth. In Section 3.6, we discuss how contending with real-world, physical constraints is also a constitutive property of robots.

[2] This expectation of traceability endures even as modern systems increasingly rely on black-box methods such as foundation models. In such cases, the exact causal chain may be too complex or inaccessible to provide a satisfying explanation, even if it exists in principle [40, 89].

don't know why I did that," or a driver makes a sudden lane change without reason. With robots, however, opacity is rarely tolerated: if an explanation cannot be produced, the system may be viewed as erroneous, untrustworthy, or unacceptably "black-boxed" [23, 150].

3.6 The Role of Embodiment

Robots occupy space, obey physical laws, and act through a material body. Unlike disembodied technology, robots cannot be separated from their embodiment: they move, collide, and gesture in ways that are constrained by mass, shape, and mechanics. This shapes user expectations—people anticipate that a robot will be visible, take up space, and exert force through its actions [29]. Embodiment also makes robot behavior more legible, since movement and presence can be directly observed [41]. At the same time, it brings expectations of safety, durability, and appropriate spatial conduct not applied to purely virtual agents [118, 129].

We do not treat embodiment as a separate "trait" but a foundational characteristic that makes the other traits legible in interaction. It renders presence physical and visible (Section 3.1): a robot is not merely conceptually "available" like software but literally occupies space, amplifying its perceived dependability and making its absence more salient. It makes sameness observable (Section 3.2): repeated motions, paths, or gestures reinforce stability, while even small variations (e.g., a wobbly gait) become salient through visible deviation. It makes responsiveness tangible (Section 3.3): a robot's physical movements (turning, grasping, rerouting) communicate instant responses in a way that purely digital systems cannot; while delays become more apparent because they are spatial and temporal. It grounds traceability in cause-and-effect perception (Section 3.5). Users can see a sensor trigger (e.g., a bump, a face detected) and watch the corresponding actuation (a turn, a wave). The physical body provides a visible chain between input and output, making the assumption of behavioral explainability more intuitive.

4 Expectations, Design Choices, and Rules

In the previous section, we set aside specific design choices to focus on the **constitutive traits** that are intrinsic to robots—qualities present even in their simplest forms of sensing and actuation. Even minimal systems, such as a light-following robot that turns toward a flashlight, a bump-and-go toy car that changes direction upon collision, or an industrial arm performing a repetitive pick-and-place routine, exhibit these traits and are therefore subject to the natural **social expectations** people hold about how robots should behave. Robots are assumed to be reliably present, consistent in behavior, immediately responsive to sensor input, act without subjective partiality, and traceable in their input-output logic, in principle.

These traits form the baseline affordances upon which we, as roboticists, design. Through **design choices**, we either reinforce or mask them. For instance, the expectation of sameness can be strengthened through repeated cues—such as a robot that always delivers a safety message in the same tone—or softened by introducing subtle variations in gesture, timing, or wording to evoke spontaneity. Similarly, responsiveness can be amplified, as when a robot vacuum immediately reroutes upon detecting an obstacle, or deliberately attenuated, as when a social robot pauses before responding to signal reflection.

HRI research systematically examines how such design variations shape experience [29, 53]. Studies have compared robots to screen-based avatars or disembodied assistants (e.g., [60, 86, 134, 140]), explored gradients of sociality from toy-like to human-like agents (e.g., [12, 73, 119]), and investigated how inherent traits can be exaggerated or masked to alter perception. Collectively, this body of work shows that reinforcing or relaxing these expectations profoundly affects the quality of interaction.

Yet, across this body of work, certain expectations and choices have congealed into **rules** that are rarely challenged. Robots are almost always designed to comply with user commands, be obedient, maintain predictability, and behave in ways that align with familiar social norms. Such rules are not formally codified but emerge as prevailing norms across robotics literature. Many studies test subtle manipulations of timing, adaptation, or personality but do so within the boundaries of these unspoken rules. Deliberately breaking them—for example, by designing robots that resist, withhold, or subvert user expectations—remains rare, yet doing so could reveal new possibilities for interaction.

5 Rules to Break

If these unspoken rules shape how robots are built and studied, then breaking them offers a way to see the field (and its assumptions) more clearly. Rather than treating these rules as fixed constraints, we approach them as design materials: norms that can be bent, inverted, or suspended to reveal novel forms of interaction. In what follows, we identify several rules that underlie HRI and explore the new opportunities that emerge when they are deliberately broken.

Rule 1: Robots Should Always Be Willing to Engage

Robotics research often features systems that are always "on," attentive, and ready to interact [55]. This ideal of continuous availability has become an tacit standard. We measure success by indicators such as sustained engagement, longer interaction time, greater up-time, improved battery performance, and fewer instances of users powering down the system or putting it away (as reviewed in [81, 93]). Then, by optimizing for these markers of success, the goal of continuous availability is thus embedded in how we design robots themselves: with eyes that continually scan the room, constant motion that signal perpetual "aliveness," never appearing to truly idle or be "off" (e.g., [6, 26, 103, 138]). Availability has been expressed through *proactive* behaviors, such as autonomously tidying up clutter in a home or navigating a facility to collect environmental data [81], and in *reactive* responses, like always responding to user queries or behavioral prompts [17]. Systems that fail to sustain this responsiveness are often seen as falling short [34, 61, 136, 141].

Continuous availability is a reasonable design norm: users never have to wait or wonder if the robot is "on" to respond [34, 101]. This fosters the trait of reliable presence (Section 3.1) and aligns with the natural expectation that a service-oriented, interactive system should always be ready and willing to engage. However, continuous availability is not always optimal—and in many real-world contexts, it can be socially inappropriate, cognitively exhausting, or simply unwelcome [136]. At times, it may be more effective for a robot to power down, "sleep," or deliberately ignore interaction attempts.

In human-human interaction, selective availability is a socially meaningful skill, where strategic non-attention reflects judgment rather than inattention. Teachers, for instance, withhold responses to off-task behavior, using silence as a classroom management tool to avoid rewarding disruption [18, 92]. Parents likewise may ignore a child's tantrum so as not to reinforce undesirable behaviors [72]. Therapists employ silence to create reflective space, signaling respect and allowing clients to process emotions [58]. Even in everyday conversations among friends or colleagues, a pause or non-response can communicate empathy, restraint, or deference to social norms [102, 146]. Rather than being a radical departure from the norm, strategic ignoring represents an established, socially meaningful practice that robots may also benefit from adopting.

In fact, we can imagine several compelling contexts where deliberate non-engagement by a robot is the socially intelligent choice. In healthcare, for example, a companion robot stationed on a hospital ward may need to ignore low-priority social bids to preserve patient rest or prioritize safety. In a classroom, a tutoring robot that answers every whispered off-task remark risks reinforcing disruption, whereas withholding a response can maintain focus for the whole group. Similarly, when users test boundaries by issuing inappropriate or repetitive prompts, selective non-responsiveness can serve as a form of behavioral shaping. Everyday interactions provide additional lessons. When speech is ambient or not directed at the robot, remaining silent may be more respectful than intruding (e.g., [51, 110, 111])—just as it would be more considerate for a robot to "look away" when someone takes a private phone call.

This discernment becomes most salient in long-term deployments, where robots live alongside users in their everyday environments. The home, for instance, is a deeply personal space: just as it would be inappropriate for a human therapist to enter someone's home uninvited and announce that it is time for therapy (even on a predefined schedule), it is likely problematic for a robot to do so [108, 110, 111]. Restraint matters in public settings as well. For example, a museum guide robot that chooses not to answer questions during a reflective art exhibit helps preserve the contemplative atmosphere that human curators intentionally create. These scenarios illustrate a broader principle: robots must be equipped not only to engage, but also to discern when non-response is the more socially intelligent action. Rather than treating always-on, continuous availability as the ideal, roboticists should consider that strategic non-responsiveness can be, at times, more socially appropriate, cognitively sustainable, and behaviorally effective.

Rule 2: Robots Should Always Offer Help

Helping is a fundamental dynamic of human-human relationships. Yet, the act of offering help can sometimes be met with resistance [15, 120]. For example, what goes on and what goes wrong when one offers to help a friend and is rudely rebuffed? It has been said that the word "help" itself comes up primarily when someone is described to have *not* been helpful [109, 120].

Robots are built to assist people. In contrast to human-human dynamics, it is generally assumed that robots should be readily available when needed and always willing to help its users [30, 42, 63]. We challenge this prevailing paradigm and can imagine situations where a robot should opt to withhold help, even when it is technically capable of assisting. For instance, in rehabilitation, withholding help can encourage independence—such as when a robot observes a user struggling slightly to stand but allows them to complete the motion on their own to support intrinsically motivated effort and learning [100]. A restaurant robot observes a waitress dropping a fork but refrains from offering help, recognizing that stepping in would interrupt the flow of professional service and draw attention. In group settings, a robot may refrain from answering to give someone a chance to recall information independently. Robots may also withhold assistance when user preferences are known (e.g., a user who prefers manual control over cooking tasks; [44, 149]) or when the context is ambiguous and premature intervention could cause confusion or offense [109]. In these cases, not helping is not a limitation of the robot, but a strategic behavior aligned with social, emotional, or pedagogical goals. This presents an opportunity for HRI research to examine how robots can discern when it is appropriate to offer or withhold assistance to users.

Rule 3: Robots Should Always Be Task-Productive

Robots are often evaluated by their efficiency and task-oriented success [128]. However, success is not always defined purely in terms of functional task performance. For instance, some studies define success as increasing the amount of eye contact users make with the robot, treating it as a proxy for engagement (e.g., [68, 79, 88, 93]). Conversational therapy systems may aim to maximize speaking time as an indicator of therapeutic progress [36, 46], while service robots often optimize for metrics such as task completion time or the number of customers served [128]. Nevertheless, these benchmarks reflect broader societal values that emphasize output, speed, and optimization in relation to functional task performance.

However, in many social and collaborative contexts, rigid adherence to task goals may inadvertently undermine relational dynamics or overlook the importance of small, seemingly "unproductive" moments that contribute to trust, rapport, and long-term acceptance. Consider, in a healthcare setting, a robot that rushes through exercises to maximize repetitions may be less effective than one that pauses to offer encouragement, even if this behavior reduces raw efficiency. In education, a tutoring robot that occasionally digresses with small talk or playful gestures may support sustained engagement better than one that delivers information as quickly as possible. In industrial contexts, robots designed only for throughput may ignore opportunities to foster camaraderie with human collaborators—for instance, by initiating light conversation on an assembly line, which could reduce monotony and stress [142].

These examples highlight a broader tension: optimization for immediate task performance can come at the expense of social value [106]. A robot that completes tasks flawlessly but leaves users stressed or alienated has failed in a deeper sense. Small, seemingly inefficient behaviors—pauses, acknowledgments, digressions—often carry disproportionate weight in human relationships. They can signal attentiveness, empathy, and respect. By ignoring these dimensions, success metrics risk flattening robots into narrow tools rather than potential partners in interaction [11].

While it may not always translate directly to immediate functional task outcomes, such interactions could foster a more positive

work environment, reduce stress, and support human well-being. Future work should reconsider what it means for a robot to be successful, expanding evaluation criteria to include social value and relational outcomes, not just efficiency. Reframing success around both productivity and social value leverages robots' duality as artifacts and as partners (Section 2), and points toward designs that sustain meaningful human-robot relationships over time.

Rule 4: Robots Should Always Be Polite and Deferential

We typically design robots to be polite (e.g., [50, 80, 84, 91, 113, 127]). Researchers incorporate system-level rules to avoid interrupting, contradicting, or confronting users, as a way of allowing the robot to signal friendliness and minimize social friction. However, politeness and deference can be counterproductive. For example, consider a rule that restricts the robot from interrupting a user while they are speaking—a generally sound and polite constraint. However, in a situation where the user begins to spiral into a repetitive or self-deprecating monologue, the rule may need to be relaxed to allow a well-timed, gentle interruption that redirects the user constructively (e.g., [108]). In this case, the rule's intent (respecting user agency) must be weighed against its current utility and possible harm.

In educational settings, a robot tutor may need to interrupt a student mid-explanation to correct a fundamental misunderstanding before it becomes entrenched. In healthcare, a robot reminding a patient about medication adherence may need to persist or escalate its tone if polite prompting is repeatedly ignored. Even in customer service, a robot may need to push back gently when a user makes an unreasonable request, such as asking it to perform actions either inappropriate or outside its scope. In these cases, assertiveness is not a breakdown in politeness but rather a demonstration of situational awareness and a commitment to supporting human goals responsibly. In sum, HRI research should explore how robots can balance politeness with assertiveness, developing context-aware strategies that allow them to interrupt, redirect, and disagree meaningfully.

Rule 5: Robots Should Never Withhold Information or Lie

It is natural to expect that robots should always provide complete and accurate information when asked. This expectation reflects a view of robots as transparent, factual tools designed to reduce uncertainty and deliver immediate answers. However, in socially and ethically complex situations, unconditional disclosure can be inappropriate or even dangerous. For example, in elder care contexts, a robot supporting a person with dementia may know where the car keys are but choose to withhold that information if there is reason to believe the person may attempt to drive unsafely or leave the house without supervision. In this case, withholding is a protective measure that prioritizes the user's physical safety over immediate compliance. Similarly, a therapeutic robot may withhold responses that could cause distress, or deliberately delay factual answers in educational contexts to promote problem-solving.

There are cases in which providing partial information is more appropriate than full disclosure. For instance, a healthcare robot might tell a patient that their test results are under review without immediately revealing abnormal findings, allowing a physician to communicate the results in a controlled clinical setting. In rare but critical situations, limited forms of strategic deception may even be ethically justified. For example, a robot might report that building exits are temporarily inaccessible during a lockdown to prevent individuals from moving toward danger. These examples demonstrate that rigid truth-telling can be socially and ethically insufficient. Future work should examine how robots can make context-sensitive decisions about when to disclose, withhold, or alter information in ways that prioritize safety, well-being, and appropriate delegation of sensitive communication.

Rule 6: Robots Should Never Make Mistakes

Robots are often designed to appear competent, consistent, and error-free—reflecting the belief that reliability and precision are core to their value as machines (Sections 3.1 and 3.2). As a result, mistakes are typically treated as design flaws to be avoided or corrected [76, 96]. However, intentional errors can serve important relational and pedagogical functions. For example, in educational settings, a robot that makes a simple mistake while solving a problem may invite the user to correct it and reinforces understanding through a "learning-by-teaching" paradigm [45]. A therapy robot designed for children with motor difficulties can occasionally drop an object or move clumsily to create opportunities for the child to "help." Such moments can foster empathy, reduce pressure for perfection, and make both the therapy and the robot itself more approachable. In another case, a robot may choose to lose at a game on purpose to boost a child's confidence or encourage continued engagement. Here, these small, human-like errors can build rapport by making the robot seem more relatable, fallible, and less intimidating. Future research should examine how robots can strategically make mistakes to support social, emotional, and learning outcomes without undermining user confidence in the system.

Rule 7: Robots Should Never Model Harmful Behavior

Robots are typically designed to avoid behaviors that are perceived as aggressive, exclusionary, or morally inappropriate—such as mocking, taunting, or bullying. Such actions are widely regarded unacceptable among humans and, by extension, are excluded from robotic conduct to maintain trust and psychological safety. However, in controlled settings, robots can strategically model norm-violating behavior to promote reflection and prosocial action. For instance, prior work has used two robots to simulate a bullying scenario, where one robot teases or excludes the other, to study how children respond as bystanders [131]. These scenarios are designed not to normalize bullying, but to prompt users to recognize mistreatment and practice appropriate intervention strategies. Another example may be a training robot for emergency response that simulates inflicting harm (e.g., applying too much force) to allow trainees to confront and analyze moral responses to machine-caused injury, without real danger. By witnessing norm violations enacted by robots, users are given a safe, repeatable context in which to explore empathy, fairness, and the moral imperative to speak up.

6 Discussion

We began this paper by outlining constitutive traits of robots. Despite the ontological ambiguity surrounding how robots are interpreted in human interaction, we argue that unique traits arise from their status as machines. Around these traits, humans intuitively build assumptions that shape how robots are perceived, how they are valued, how they are designed, and how HRI research is conducted. The rules we later propose for breaking reflect prevailing norms in our research community. We argue that these norms should be challenged, since breaking them may in some cases lead to more socially productive outcomes.

One might argue that a "rule with exceptions" is best understood as a single, more complex rule, rather than a rule plus its violation. However, the pattern of HRI literature demonstrates that, though these rules are not rigid or explicitly codified, they are generally respected and rarely defied. Throughout this paper, we referenced relevant reviews and studies to convey the broader acceptance of these assumptions. Therefore, our motivation for this paper is more to challenge the prevailing norms of HRI research than simply advocate designing robots that follow complex or conditional rules.

6.1 Further Betraying Robotic Traits

Our two lists—the constitutive traits (Section 3) and the rules to break (Section 5)—are not exhaustive or mutually exclusive. Rather, they are broad sketches: one highlights expectations intrinsic to robots, while the other calls attention to research norms that merit questioning. Although our paper does not directly map each constitutive trait to each rule, their connections are apparent and overlapping in several cases. For instance, Rule 1 (*Robots Should Always Be Willing to Engage*) arises from robots' *Reliability of Presence*. Similarly, Rule 3 (*Robots Should Always Be Task Productive*) can be traced to the traits of *Sameness and Precision* and *Responsiveness*: robots' precision and consistency lead us to value their functional utility, judge their success by task completion, and treat sociability as a trade-off against efficiency. Accordingly, Rule 6 (*Robots Should Never Make Mistakes*) also stems from the expectation of *Sameness*.

Given this, we can explore how to design robots that subvert their intrinsic traits to yield socially productive outcomes. For example, we can subvert the trait of *Sameness* into an idea of generative inconsistency. Since robots are rapidly deployed into personal spaces for longer-term use, we can imagine robots that live and thus "grow" alongside us. Robots that age, degrade, or evolve, showing wear not as a mechanical failure but as identity. Practically, this might take the form of a robot for children that begins with a higher-pitched voice and limited vocabulary, then gradually "matures" into an adult-like voice with richer prosody and more advanced communicative reasoning as the user enters adulthood. Similarly, a robot designed for young families might initially engage in energetic, playful group interactions, then gradually adopt calmer, more focused dyadic behaviors as the children grow older. Here, the robot retains its social utility by disrupting assumptions related to *Sameness*, allowing it to remain relevant rather than outgrown.

Across constitutive traits, robots are typically designed to minimize awkwardness: they respond quickly, act consistently, remain neutral, embody space predictably, and withdraw smoothly when tasks are complete. Purposefully designing awkward behaviors destabilizes these assumptions in subtle but socially meaningful ways. For instance, it may be counterintuitive to design robots that stand too close or face the wrong direction. However, we can imagine a conference robot that deliberately positions itself in a "bad" spot to encourage attendees to physically rearrange and interact with one another. Exploring how to design such interactions could reveal new insights into the spatial politics of bodies, group dynamics, and social bonding through collective negotiation. Similarly, in therapy for individuals with memory impairments, a companion robot could use "playful incongruity" by sharing implausible or incorrect anecdotes. This behavior can stimulate memory recall and reasoning by prompting users to detect what feels "out of place." In these cases, awkwardness becomes a resource to reconfigure social interaction, invite reflection, and deepen cognitive engagement.

6.2 Breaking the "More Human-Like" Rule

So far, we have described rule-breaking mainly as a departure from robots' machine-like traits—adding variation, hesitation, or defiance to soften their mechanical character. In this view, rule-breaking serves to make robots feel more *human-like*. This perspective is intuitive: it reflects the observation that humans routinely bend or break social norms; robots that do so may therefore appear more authentic, socially attuned, and relatable.

However, breaking rules need not always mean humanizing robots. In some contexts, the more meaningful "rule break" is to resist the pull of over-socialization and instead preserve the very machine-like traits that are sometimes treated as limitations. A robot that insists on repeating a warning in the same tone, that acts without hesitation, or that maintains rigid impartiality may seem overly mechanical in casual social interaction, but in domains like surgery, aviation, or industrial safety, those same qualities are what inspire confidence and trust. Here, these traits become the virtues of robots, not available with humans (Section 3).

Acknowledging this duality helps clarify the scope of our argument. Rule-breaking is not a unidirectional path toward greater human-likeness, nor is it a blanket rejection of the traits that define machines. Instead, it can take two forms: (1) breaking rules to align more closely with human social practice, or (2) breaking rules by reaffirming and amplifying machine-like qualities in contexts where those qualities are socially or functionally valuable.

Furthermore, this raises the question of whether a robot can become "more social" without becoming "more human-like." These concepts are often conflated in HRI research, although they do not always need to align. We can imagine how a robot's sociality can derive, not from imitation of human traits, but from its capacity to behave in ways that humans *cannot* or *would not* enact themselves.

Consider a robot embedded in a workplace or classroom over several years. By tracking interaction patterns (who collaborates, who is left out), it could reorganize groups or suggest new partnerships. Humans rarely sustain impartial, long-term social memory at such scale, yet this form of mediation could cultivate more equitable participation. In a brainstorming session, people often converge prematurely on one idea [64, 90]. A robot, by contrast, could introduce a deliberately contradictory or improbable suggestion based on lateral semantic association—something no person would seriously propose. Though nonhuman in logic, such contributions

might provoke novel connections and richer group discussion. In shared environments, collective stress or fatigue often accumulates invisibly. An ambient robot could monitor environmental cues (noise, tone, activity) and adjust lighting, sound, or scent to gently regulate the collective mood. Acting as a social barometer, it represents group-level emotional states that individuals intuitively sense but cannot express in real time. Across these cases, robots function as agents of social *facilitation* rather than social *imitation*, producing new, nonhuman forms of social intelligence.

Lastly, we did not treat embodiment as a distinct trait, as it is core to what defines a robot and is also shared with humans. Yet, embodiment critically mediates the expression of all other traits (Section 3.6), and can manifest in uniquely robotic ways. Robots are often imagined as cold, rigid, or plastic, yet they could be designed to subvert this material expectation in ways that generate new forms of social value. For example, a social robot might expand or inflate to become physically larger in group settings—drawing attention to quieter participants, redistributing focus, or signaling a need for collective attention. Similarly, a therapy robot could vary its material properties, becoming soft, warm, and pliable during moments of comfort, then hardening when idle or unavailable outside of the therapy context (e.g., expressing a clear break of Rule 1). Such shifts could provide tactile cues of safety and care during interaction while maintaining clear boundaries beyond it.

6.3 Ethics Prohibitions

In this paper, we used "rules" in a broad sense to encompasses seven examples, but it is useful to distinguish several layers that reveal different kinds of rules robots may break. *Constitutive traits* are the inherent machine-like properties of robots (e.g., sameness, responsiveness, traceability) that give rise to user expectations. *Social expectations* are the intuitive assumptions people hold about robots based on those traits (e.g., that robots will not hesitate or defy user commands). *Design norms* are the engineering conventions that codify these expectations into practice (e.g., building systems that always respond or follow predictable motion paths; else, departures from such standards are viewed as design flaws or system errors). Together, these three layers define a practical space where rule-breaking in HRI can occur. Our interest lies not merely in *whether* robots should break such rules, but in *which* rules they might productively violate, *when*, and *for what purpose*.

Beyond these, however, lies a distinct category of *ethical prohibitions*: the conceptual rules that constrain robots absolutely (e.g., never harm, never deceive, never override human agency). These prohibitions define the moral limits within which productive rule-breaking can occur. Recognizing this distinction clarifies that our argument is not for unbounded disobedience, but for thoughtful design strategies that balance creativity with ethical constraint. Rule-breaking can occur at any of these four layers, but with different implications: playful or deliberate breaks may be productive at the levels of expectation or design, whereas violating ethical prohibitions risks undermining safety and trust.

In ethics, truly absolute prohibitions are almost always negative (*never* do X), rather than positive mandates (*always* do Y). For robotics, this asymmetry is important. Negative rules do not dictate what the robot should do in every situation, but they sharply delimit the range of acceptable actions. In this sense, negative absolutes reduce the action space without fully determining it. By contrast, attempting to impose absolute positive rules on robots—for example, "always tell the truth" or "always act in the user's best interest"—introduces ambiguity. What counts as truth? How do we define the user's "best interest" in cases of conflicting needs or goals? Positive mandates require contextual interpretation and value judgments, which cannot be captured by exceptionless formulations.

This framing suggests that"absolute rules" robots should "never break" will rarely be prescriptive. Instead, they will be protective constraints: ethical boundaries that ensure safety, transparency, and respect for autonomy, while leaving space for flexibility, adaptation, and even productive rule-breaking within those limits. Acknowledging where rules must remain inviolable does not weaken the case for rule-breaking; it strengthens it by clarifying where responsible experimentation can and should occur.

Against this backdrop, there are several widely recognized non-negotiables in HRI [7, 148]. Chief among them is the imperative that robots must never cause physical harm to humans. This principle is reminiscent of Asimov's First Law but also grounded in real-world engineering ethics and legal frameworks. A robot that directly injures people undermines the baseline of safety upon which all other forms of trust are built. Even playful or socially enriching behaviors lose meaning if physical safety cannot be assumed.

Building on this foundation, we propose an additional non-negotiable: a robot should never conceal its purpose. Mystery about *how* a robot functions (opacity about its sensors, algorithms, or control systems) can be acceptable, even desirable in contexts like art or entertainment. However, mystery about *why* a robot behaves in a certain way (i.e., what its goals are, who it serves, and what outcomes it seeks) quickly undermines trust. Users must be able to discern a robot's purpose in order to make informed decisions about how to interact with it, what boundaries to set, and whether or not to rely on it. Transparency of intent is thus not a design courtesy, but a prerequisite for ethical and trustworthy HRI.

7 Conclusion

Robots have long been framed as ideal rule-followers, but we argue that strict obedience can limit their potential to support human social good. By outlining prevailing norms in HRI and showing how deliberately breaking them can yield more effective or morally attuned outcomes, we suggest that rule-breaking can be a purposeful design strategy. Embracing the idea of robots that resist, deceive, or err in contextually appropriate ways allows designers to create systems that engage more authentically with humans and reflect the complex, situational ethics that govern social life itself.

Acknowledgments

This work was supported by the National Science Foundation (NSF) under Grant IIS-2106690 and an *Envisioning Artificial Intelligence at Yale* seed grant. R. Ramnauth is supported by the NSF GRFP and the National Academies of Sciences, Engineering, and Medicine (NASEM) Ford Fellowships. The opinions and recommendations expressed in this material are those of the authors and do not necessarily reflect the views of these funding agencies.

References

[1] Jamil Abou Saleh and Fakhreddine Karray. 2010. Towards generalized performance metrics for human-robot interaction. In *2010 International Conference on Autonomous and Intelligent Systems, AIS 2010*. IEEE, 1–6.

[2] Henny Admoni, Anca Dragan, Siddhartha S Srinivasa, and Brian Scassellati. 2014. Deliberate delays during robot-to-human handovers improve compliance with gaze communication. In *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction*. 49–56.

[3] Safinah Ali, Nisha Devasia, Hae Won Park, and Cynthia Breazeal. 2021. Social robots as creativity eliciting agents. *Frontiers in Robotics and AI* 8 (2021), 673730.

[4] Safinah Ali, Tyler Moroso, and Cynthia Breazeal. 2019. Can children learn creativity from a social robot?. In *Proceedings of the 2019 Conference on Creativity and Cognition*. 359–368.

[5] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. 2016. Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565* (2016).

[6] Kika Arias, Sooyeon Jeong, Hae Won Park, and Cynthia Breazeal. 2020. Toward designing user-centered idle behaviors for social robots in the home. In *1st international workshop on Designerly HRI Knowledge. Held in conjunction with the 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN 2020)*.

[7] Thomas Arnold and Matthias Scheutz. 2017. Beyond moral dilemmas: exploring the ethical landscape in HRI. In *Proceedings of the 2017 ACM/IEEE international conference on human-robot interaction*. 445–452.

[8] Thomas Arnold and Matthias Scheutz. 2020. HRI ethics and type-token ambiguity: what kind of robotic identity is most responsible? *Ethics and Information Technology* 22, 4 (2020), 357–366.

[9] Isaac Asimov. 1942. Runaround. *Astounding Science Fiction* 29, 3 (March 1942), 94–103. First appearance of the "Three Laws of Robotics".

[10] Isaac Asimov. 2004. *I, robot*. Vol. 1. Spectra.

[11] Aorigele Bao, Yi Zeng, and Enmeng Lu. 2023. Mitigating emotional risks in human-social robot interactions through virtual interactive environment indication. *Humanities and Social Sciences Communications* 10, 1 (2023), 1–9.

[12] Olivia Barber, Eszter Somogyi, E Anne McBride, and Leanne Proops. 2023. Exploring the role of aliveness in children's responses to a dog, biomimetic robot, and toy dog. *Computers in Human Behavior* 142 (2023), 107660.

[13] Christoph Bartneck, Chioke Rosalia, Rutger Menges, and Inèz Deckers. 2005. Robot Abuse—A Limitation of the Media Equation. (2005).

[14] Jenay M Beer, Arthur D Fisk, and Wendy A Rogers. 2014. Toward a framework for levels of robot autonomy in human-robot interaction. *Journal of human-robot interaction* 3, 2 (2014), 74.

[15] Harry Beilin. 2013. *The Development of Prosocial Behavior*. Academic Press.

[16] Cindy L Bethel, Matthew R Stevenson, and Brian Scassellati. 2011. Secret-sharing: Interactions between a child, robot, and adult. In *2011 IEEE International Conference on systems, man, and cybernetics*. IEEE, 2489–2494.

[17] Lauriane Blavette, Sébastien Dacunha, Xavier Alameda-Pineda, Daniel Hernández García, Sharon Gannot, Florian Gras, Nancie Gunson, Séverin Lemaignan, Michal Polic, Pinchas Tandeitnik, et al. 2025. Acceptability and Usability of a Socially Assistive Robot Integrated With a Large Language Model for Enhanced Human-Robot Interaction in a Geriatric Care Institution: Mixed Methods Evaluation. *JMIR Human Factors* 12, 1 (2025), e76496.

[18] Iris EWM Bogaers, Jan Willem Grijpma, Marjolein K Camphuijsen, Anne de la Croix, and Chiel van der Veen. 2025. Different voices of silence: a Q methodology study of teachers' perceptions on silence in the classroom. *Teaching and Teacher Education* 165 (2025), 105136.

[19] Valentino Braitenberg. 1986. *Vehicles: Experiments in synthetic psychology*. MIT press.

[20] Cynthia Breazeal. 2004. Social interactions in HRI: the robot view. *IEEE transactions on systems, man, and cybernetics, part C (applications and reviews)* 34, 2 (2004), 181–186.

[21] Cynthia Breazeal, Kerstin Dautenhahn, and Takayuki Kanda. 2016. Social robotics. *Springer handbook of robotics* (2016), 1935–1972.

[22] Elizabeth Broadbent, Rebecca Stafford, and Bruce MacDonald. 2009. Acceptance of healthcare robots for the older population: Review and future directions. *International journal of social robotics* 1, 4 (2009), 319–330.

[23] Jenna Burrell. 2016. How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big data & society* 3, 1 (2016), 2053951715622512.

[24] Julie Carpenter. 2016. *Culture and human-robot interaction in militarized spaces: A war story*. Routledge.

[25] Houston Claure, Mai Lee Chang, Seyun Kim, Daniel Omeiza, Martim Brandao, Min Kyung Lee, and Malte Jung. 2022. Fairness and transparency in human-robot interaction. In *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 1244–1246.

[26] Raymond H Cuijpers and Marco AMH Knops. 2015. Motions of robots matter! the social effects of idle and meaningful motions. In *International conference on social robotics*. Springer, 174–183.

[27] Kate Darling. 2016. Extending legal protection to social robots: The effects of anthropomorphism, empathy, and violent behavior towards robotic objects. In *Robot Law*. Edward Elgar Publishing, 213–232.

[28] Kate Darling. 2020. Why people become strangely attached to their robot vacuum cleaners. *New Scientist* (18 March 2020). https://www.newscientist.com/article/mg24532741-700-why-people-become-strangely-attached-to-their-robot-vacuum-cleaners/

[29] Kerstin Dautenhahn. 2007. Socially intelligent robots: dimensions of human–robot interaction. *Philosophical transactions of the royal society B: Biological sciences* 362, 1480 (2007), 679–704.

[30] Kerstin Dautenhahn, Sarah Woods, Christina Kaouri, Michael L Walters, Kheng Lee Koay, and Iain Werry. 2005. What is a robot companion—friend, assistant or butler?. In *2005 IEEE/RSJ International Conference on Intelligent Robots and Systems*. 1192–1197. doi:10.1109/IROS.2005.1545189

[31] Maartje De Graaf, Somaya Ben Allouch, and Jan Van Dijk. 2017. Why do they refuse to use my robot? Reasons for non-use derived from a long-term home study. In *Proceedings of the 2017 ACM/IEEE international conference on human-robot interaction*. 224–233.

[32] Maartje Ma De Graaf, Somaya Ben Allouch, and Tineke Klamer. 2015. Sharing a life with Harvey: Exploring the acceptance of and relationship-building with a social robot. *Computers in human behavior* 43 (2015), 1–14.

[33] Daniel C Dennett. 1971. Intentional systems. *The journal of philosophy* 68, 4 (1971), 87–106.

[34] Amol Deshmukh, Katrin Solveig Lohan, Gnanathusharan Rajendran, and Ruth Aylett. 2018. social impact of recharging activity in long-Term hri and Verbal strategies to Manage User expectations During recharge. *Frontiers in Robotics and AI* 5 (2018), 23.

[35] Andrea Deublein, Anne Pfeifer, Katinka Merbach, Katharina Bruckner, Christoph Mengelkamp, and Birgit Lugrin. 2018. Scaffolding of motivation in learning using a social robot. *Computers & Education* 125 (2018), 182–190.

[36] Hang Ding, Joshua Simmich, Atiyeh Vaezipour, Nicole Andrews, and Trevor Russell. 2024. Evaluation framework for conversational agents with artificial intelligence in health interventions: a systematic scoping review. *Journal of the American Medical Informatics Association* 31, 3 (2024), 746–761.

[37] Francesca Dino, Rohola Zandie, Hojjat Abdollahi, Sarah Schoeder, and Mohammad H Mahoor. 2019. Delivering cognitive behavioral therapy using a conversational social robot. In *2019 IEEE/RSJ international conference on intelligent robots and systems (IROS)*. IEEE, 2089–2095.

[38] Louis DiPietro. 2024. Death of a robot: Repairing robots fosters social bonds. *Cornell Chronicle* (21 Oct. 2024). https://news.cornell.edu/stories/2024/10/death-robot-repairing-robots-fosters-social-bonds

[39] Thomas Dohmen and Jan Sauermann. 2016. Referee bias. *Journal of Economic Surveys* 30, 4 (2016), 679–695.

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

[41] Anca D Dragan, Kenton CT Lee, and Siddhartha S Srinivasa. 2013. Legibility and predictability of robot motion. In *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 301–308.

[42] Neta Ezer, Arthur D Fisk, and Wendy A Rogers. 2009. Attitudinal and intentional acceptance of domestic robots by younger and older adults. In *Universal Access in Human-Computer Interaction. Intelligent and Ubiquitous Interaction Environments: 5th International Conference, UAHCI 2009, Held as Part of HCI International 2009, San Diego, CA, USA, July 19-24, 2009. Proceedings, Part II 5*. Springer, 39–48.

[43] Davide Falanga, Suseong Kim, and Davide Scaramuzza. 2019. How fast is too fast? the role of perception latency in high-speed sense and avoid. *IEEE Robotics and Automation Letters* 4, 2 (2019), 1884–1891.

[44] Matteo Meregalli Falerni, Vincenzo Pomponi, Hamid Reza Karimi, Matteo Lavit Nicora, Le Anh Dao, Matteo Malosio, and Loris Roveda. 2024. A framework for human–robot collaboration enhanced by preference learning and ergonomics. *Robotics and Computer-Integrated Manufacturing* 89 (2024), 102781.

[45] Logan Fiorella and Richard E Mayer. 2013. The relative benefits of learning by teaching and teaching expectancy. *Contemporary Educational Psychology* 38, 4 (2013), 281–288.

[46] Nikolaos Flemotomos, Victor R Martinez, Zhuohao Chen, Karan Singla, Victor Ardulov, Raghuveer Peri, Derek D Caperton, James Gibson, Michael J Tanana, Panayiotis Georgiou, et al. 2022. Automated evaluation of psychotherapy skills using speech and language technologies. *Behavior Research Methods* 54, 2 (2022), 690–711.

[47] Terrence Fong, Illah Nourbakhsh, and Kerstin Dautenhahn. 2003. A survey of socially interactive robots. *Robotics and autonomous systems* 42, 3-4 (2003), 143–166.

[48] Jodi Forlizzi and Carl DiSalvo. 2006. Service robots in the domestic environment: a study of the roomba vacuum in the home. In *Proceedings of the 1st ACM SIGCHI/SIGART conference on Human-robot interaction*. 258–265.

[49] Iason Gabriel. 2020. Artificial intelligence, values, and alignment. *Minds and machines* 30, 3 (2020), 411–437.

[50] Elisabeth Ganal, Michelle Habenicht, and Birgit Lugrin. 2024. Excuse Me, May I Disturb You? The Influence of Politeness of a Social Robot on the Perception of Interruptions. In *2024 33rd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2240–2247.

[51] Nicholas C Georgiou, Rebecca Ramnauth, Emmanuel Adeniran, Michael Lee, Lila Selin, and Brian Scassellati. 2023. Is someone there or is that the tv? detecting social presence using sound. *ACM Transactions on Human-Robot Interaction* 12, 4 (2023), 1–33.

[52] Mojgan Ghanbari, Shahed Rasekh, Amaro Fernandes de Sousa, and Pedro Fonseca. 2024. Towards the always-on operation of mobile service robots. In *2024 7th Iberian Robotics Conference (ROBOT)*. IEEE, 1–6.

[53] Michael A Goodrich, Alan C Schultz, et al. 2008. Human–robot interaction: a survey. *Foundations and trends® in human–computer interaction* 1, 3 (2008), 203–275.

[54] Markus Gsell and Peter Gomber. 2009. Algorithmic trading engines versus human traders-Do they behave different in securities markets? (2009).

[55] Nick Hawes, Christopher Burbridge, Ferdian Jovan, Lars Kunze, Bruno Lacerda, Lenka Mudrova, Jay Young, Jeremy Wyatt, Denise Hebesberger, Tobias Kortner, et al. 2017. The strands project: Long-term autonomy in everyday environments. *IEEE Robotics & Automation Magazine* 24, 3 (2017), 146–156.

[56] Frank Hegel, Claudia Muhl, Britta Wrede, Martina Hielscher-Fastabend, and Gerhard Sagerer. 2009. Understanding social robots. In *2009 Second International Conferences on Advances in Computer-Human Interactions*. IEEE, 169–174.

[57] Anna Henschel, Guy Laban, and Emily S Cross. 2021. What makes a robot social? A review of social robots from science fiction to a home or hospital near you. *Current Robotics Reports* 2, 1 (2021), 9–19.

[58] Clara E Hill, Barbara J Thompson, and Nicholas Ladany. 2003. Therapist use of silence in therapy: A survey. *Journal of clinical psychology* 59, 4 (2003), 513–524.

[59] Tom Hitron, Noa Morag Yaar, and Hadas Erel. 2023. Implications of ai bias in hri: Risks (and opportunities) when interacting with a biased robot. In *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*. 83–92.

[60] Thomas Holz, Mauro Dragone, and Gregory MP O'Hare. 2009. Where robots and virtual agents meet: a survey of social interaction research across milgram's reality-virtuality continuum. *International Journal of Social Robotics* 1, 1 (2009), 83–93.

[61] Shanee Honig and Tal Oron-Gilad. 2018. Understanding and resolving failures in human-robot interaction: Literature review and model development. *Frontiers in psychology* 9 (2018), 861.

[62] Rowan Hooper. 2019. When robots are ultra-lifelike, will it be murder to switch one off? *New Scientist* (20 Nov. 2019). https://www.newscientist.com/article/2223763-when-robots-are-ultra-lifelike-will-it-be-murder-to-switch-one-off/

[63] Aike C Horstmann and Nicole C Krämer. 2019. Great expectations? Relation of previous experiences with social robots in real life or in the media and expectancies based on qualitative and quantitative assessment. *Frontiers in Psychology* 10 (2019), 939.

[64] Irving L Janis. 2008. Groupthink. *IEEE Engineering Management Review* 36, 1 (2008), 36.

[65] Hifza Javed and Chung Hyuk Park. 2019. Interactions with an empathetic agent: Regulating emotions and improving engagement in autism. *IEEE robotics & automation magazine* 26, 2 (2019), 40–48.

[66] Raya A Jones. 2017. What makes a robot 'social'? *Social studies of science* 47, 4 (2017), 556–579.

[67] Michiel Joosse, Manja Lohse, Jorge Gallego Perez, and Vanessa Evers. 2013. What you do is who you are: The role of task context in perceived social robot personality. In *2013 IEEE International Conference on Robotics and Automation*. IEEE, 2134–2139.

[68] Magnus Jung, Ahmed Abdelrahman, Thorsten Hempel, Basheer Al-Tawil, Qiaoyue Yang, Sven Wachsmuth, and Ayoub Al-Hamadi. 2025. Eye contact based engagement prediction for efficient human–robot interaction. *Complex & Intelligent Systems* 11, 7 (2025), 286.

[69] Peter H Kahn Jr, Heather E Gary, and Solace Shen. 2013. Children's social relationships with current and near-future robots. *Child Development Perspectives* 7, 1 (2013), 32–37.

[70] Takayuki Kanda, Michihiro Shimada, and Satoshi Koizumi. 2012. Children learning with a social robot. In *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction*. 351–358.

[71] Takayuki Kanda, Masahiro Shiomi, Zenta Miyashita, Hiroshi Ishiguro, and Norihiro Hagita. 2009. An affective guide robot in a shopping mall. In *Proceedings of the 4th ACM/IEEE international conference on Human robot interaction*. 173–180.

[72] Michael G Kavan, Shailendra K Saxena, and Naureen Rafiq. 2018. General parenting strategies: Practical suggestions for common child behavior issues. *American family physician* 97, 10 (2018), 642–648.

[73] Megan S Kelley, J Adam Noah, Xian Zhang, Brian Scassellati, and Joy Hirsch. 2021. Comparison of human social brain activity during eye-contact with another human and a humanoid robot. *Frontiers in Robotics and AI* 7 (2021), 599581.

[74] Rucha Khot, Minha Lee, Alexandra Bejarano, Lux Miranda, Gisela Reyes-Cruz, Joel E Fischer, and Dimosthenis Kontogiorgos. 2024. Robo-Identity: Designing for Identity in the Shared World. In *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*. 1326–1328.

[75] Cory D Kidd and Cynthia Breazeal. 2008. Robots at home: Understanding long-term human-robot interaction. In *2008 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 3230–3235.

[76] Su Kyoung Kim, Elsa Andrea Kirchner, Lukas Schloßmüller, and Frank Kirchner. 2020. Errors in human-robot interactions and their effects on robot learning. *Frontiers in Robotics and AI* 7 (2020), 558531.

[77] Elena Knox and Katsumi Watanabe. 2018. AIBO robot mortuary rites in the Japanese cultural context. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020–2025.

[78] Bing Cai Kok and Harold Soh. 2020. Trust in robots: Challenges and opportunities. *Current robotics reports* 1, 4 (2020), 297–309.

[79] Kyveli Kompatsiari, Francesco Bossi, and Agnieszka Wykowska. 2021. Eye contact during joint attention with a humanoid robot modulates oscillatory brain activity. *Social cognitive and affective neuroscience* 16, 4 (2021), 383–392.

[80] Shikhar Kumar, Eliran Itzhak, Yael Edan, Galit Nimrod, Vardit Sarne-Fleischmann, and Noam Tractinsky. 2022. Politeness in human–robot interaction: a multi-experiment study with non-humanoid robots. *International Journal of Social Robotics* 14, 8 (2022), 1805–1820.

[81] Lars Kunze, Nick Hawes, Tom Duckett, Marc Hanheide, and Tomáš Krajník. 2018. Artificial intelligence for long-term robot autonomy: A survey. *IEEE Robotics and Automation Letters* 3, 4 (2018), 4023–4030.

[82] Przemyslaw A Lasota, Terrence Fong, Julie A Shah, et al. 2017. A survey of methods for safe human-robot interaction. *Foundations and Trends® in Robotics* 5, 4 (2017), 261–349.

[83] Kwan Min Lee, Wei Peng, Seung-A Jin, and Chang Yan. 2006. Can robots manifest personality?: An empirical test of personality recognition, social responses, and social presence in human–robot interaction. *Journal of communication* 56, 4 (2006), 754–772.

[84] Namyeon Lee, Jeonghun Kim, Eunji Kim, and Ohbyung Kwon. 2017. The influence of politeness behavior on user compliance with social robots in a healthcare service setting. *International journal of social robotics* 9, 5 (2017), 727–743.

[85] Iolanda Leite, André Pereira, Ginevra Castellano, Samuel Mascarenhas, Carlos Martinho, and Ana Paiva. 2011. Modelling empathy in social robotic companions. In *International conference on user modeling, adaptation, and personalization*. Springer, 135–147.

[86] Daniel Leyzberg, Samuel Spaulding, Mariya Toneva, and Brian Scassellati. 2012. The physical presence of a robot tutor increases cognitive learning gains. In *Proceedings of the annual meeting of the cognitive science society*, Vol. 34.

[87] Karolis Liaudinskas. 2022. *Human vs. machine: Disposition effect among algorithmic and human day traders*. Number 6/2022. Working Paper.

[88] Samuli Linnunsalo, Dennis Küster, Santeri Yrttiaho, Mikko J Peltola, and Jari K Hietanen. 2023. Psychophysiological responses to eye contact with a humanoid robot: Impact of perceived intentionality. *Neuropsychologia* 189 (2023), 108668.

[89] Zachary C Lipton. 2018. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue* 16, 3 (2018), 31–57.

[90] Albert J Lott and Bernice E Lott. 1961. Group cohesiveness, communication level, and conformity. *The Journal of Abnormal and Social Psychology* 62, 2 (1961), 408.

[91] Eleonore Lumer and Hendrik Buschmeier. 2023. Should robots be polite? Expectations about politeness in human–robot interaction. *Frontiers in Robotics and AI* 10 (2023), 1242127.

[92] Charles H Madsen Jr, Wesley C Becker, and Don R Thomas. 1968. Rules, praise, and ignoring: elements of elementary classroom control 1. *Journal of applied behavior analysis* 1, 2 (1968), 139–150.

[93] Kayla Matheus, Rebecca Ramnauth, Brian Scassellati, and Nicole Salomons. 2025. Long-Term Interactions with Social Robots: Trends, Insights, and Recommendations. *ACM Transactions on Human-Robot Interaction* 14, 3 (2025), 1–42.

[94] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)* 54, 6 (2021), 1–35.

[95] Lux Miranda, Ginevra Castellano, and Katie Winkle. 2023. Examining the state of robot identity. In *Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*. 658–662.

[96] Nicole Mirnig, Gerald Stollnberger, Markus Miksch, Susanne Stadler, Manuel Giuliani, and Manfred Tscheligi. 2017. To err is robot: How humans assess and act toward an erroneous social robot. *Frontiers in Robotics and AI* 4 (2017), 21.

[97] Bilge Mutlu, Fumitaka Yamaoka, Takayuki Kanda, Hiroshi Ishiguro, and Norihiro Hagita. 2009. Nonverbal leakage in robots: communication of intentions through seemingly unintentional behavior. In *Proceedings of the 4th ACM/IEEE international conference on Human robot interaction*. 69–76.

[98] Clifford Nass and Youngme Moon. 2000. Machines and mindlessness: Social responses to computers. *Journal of social issues* 56, 1 (2000), 81–103.

[99] Clifford Nass, Jonathan Steuer, and Ellen R Tauber. 1994. Computers are social actors. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 72–78.

[100] Benjamin A Newman, Reuben M Aronson, Kris Kitani, and Henny Admoni. 2022. Helping people through space and time: Assistance as a perspective on human-robot interaction. *Frontiers in Robotics and AI* 8 (2022), 720319.

[101] Gennaro Notomista and Magnus Egerstedt. 2020. Persistification of robotic tasks. *IEEE Transactions on Control Systems Technology* 29, 2 (2020), 756–767.

[102] Michael R Parke, Subrahmaniam Tangirala, Apurva Sanaria, and Srinivas Ekkirala. 2022. How strategic silence enables employee voice to be valued and rewarded. *Organizational Behavior and Human Decision Processes* 173 (2022), 104187.

[103] Hannah Pelikan and Emily Hofstetter. 2023. Managing delays in human-robot interaction. *ACM Transactions on Computer-Human Interaction* 30, 4 (2023), 1–42.

[104] Roberto Pinillos, Samuel Marcos, Raul Feliz, Eduardo Zalama, and Jaime Gómez-García-Bermejo. 2016. Long-term assessment of a service robot in a hotel environment. *Robotics and Autonomous Systems* 79 (2016), 40–57.

[105] Giovanni Pioggia, ML Sica, Marcello Ferro, Roberta Igliozzi, Filippo Muratori, Arti Ahluwalia, and Danilo De Rossi. 2007. Human-robot interaction in autism: FACE, an android-based social therapy. In *RO-MAN 2007-the 16th IEEE international symposium on robot and human interactive communication*. IEEE, 605–612.

[106] Tony J Prescott and Julie M Robillard. 2021. Are friends electric? The benefits and risks of human-robot relationships. *Iscience* 24, 1 (2021).

[107] Aditi Ramachandran, Chien-Ming Huang, and Brian Scassellati. 2019. Toward effective robot–child tutoring: Internal motivation, behavioral intervention, and learning outcomes. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 9, 1 (2019), 1–23.

[108] Rebecca Ramnauth, Emmanuel Adéníran, Timothy Adamson, Michal A Lewkowicz, Rohit Giridharan, Caroline Reiner, and Brian Scassellati. 2022. A Social Robot for Improving Interruptions Tolerance and Employability in Adults with ASD. In *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 4–13.

[109] Rebecca Ramnauth, Dražen Brščić, and Brian Scassellati. 2024. Should I Help?: A Skill-Based Framework for Deciding Socially Appropriate Assistance in Human-Robot Interactions. In *2024 33rd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2051–2058.

[110] Rebecca Ramnauth, Drazen Brščić, and Brian Scassellati. 2025. From Fidgeting to Focused: Developing Robot-Enhanced Social-Emotional Therapy (RESET) for School De-Escalation Rooms. In *Proceedings of the 34th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*.

[111] Rebecca Ramnauth, Drazen Brščić, and Brian Scassellati. 2025. A Robot-Assisted Approach to Small Talk Training for Adults with ASD. In *Proceedings of the Robotics: Science and Systems Conference (RSS)*.

[112] Byron Reeves and Clifford Nass. 1996. The media equation: How people treat computers, television, and new media like real people. *Cambridge, UK* 10, 10 (1996), 19–36.

[113] Patrizia Ribino. 2023. The role of politeness in human–machine interactions: a systematic literature review and future perspectives. *Artificial Intelligence Review* 56, Suppl 1 (2023), 445–482.

[114] Lionel P Robert Jr, Rasha Alahmad, Connor Esterwood, Sangmi Kim, Sangseok You, Qiaoning Zhang, et al. 2020. A review of personality in human–robot interactions. *Foundations and Trends® in Information Systems* 4, 2 (2020), 107–212.

[115] Ben Robins, Kerstin Dautenhahn, R Te Boekhorst, and Aude Billard. 2005. Robotic assistants in therapy and education of children with autism: can a small humanoid robot help encourage social interaction skills? *Universal access in the information society* 4, 2 (2005), 105–120.

[116] Astrid M Rosenthal-von der Pütten, Nicole C Krämer, Laura Hoffmann, Sabrina Sobieraj, and Sabrina C Eimler. 2013. An experimental study on emotional reactions towards a robot. *International Journal of Social Robotics* 5, 1 (2013), 17–34.

[117] Stuart Russell. 2022. Human-Compatible Artificial Intelligence. *Human-like machine intelligence* 1 (2022), 3–22.

[118] Selma Šabanović. 2010. Robots in society, society in robots: Mutual shaping of society and technology as a framework for social robot design. *International Journal of Social Robotics* 2, 4 (2010), 439–450.

[119] Brian Scassellati, Henny Admoni, and Maja Matarić. 2012. Robots for use in autism research. *Annual review of biomedical engineering* 14, 1 (2012), 275–294.

[120] Edgar H. Schein. 2009. *Helping: How to Offer, Give, and Receive Help*. Berrett-Koehler Pub., San Francisco.

[121] John R Searle. 1980. Minds, brains, and programs. *Behavioral and brain sciences* 3, 3 (1980), 417–424.

[122] Danish Shaikh and Ignacio Rañó. 2020. Braitenberg vehicles as computational tools for research in neuroscience. *Frontiers in bioengineering and biotechnology* 8 (2020), 565963.

[123] Thomas B Sheridan. 2016. Human–robot interaction: status and challenges. *Human factors* 58, 4 (2016), 525–532.

[124] Takanori Shibata and Kazuyoshi Wada. 2011. Robot therapy: a new approach for mental healthcare of the elderly–a mini-review. *Gerontology* 57, 4 (2011), 378–386.

[125] Toshiyuki Shiwa, Takayuki Kanda, Michita Imai, Hiroshi Ishiguro, and Norihiro Hagita. 2009. How quickly should a communication robot respond? Delaying strategies and habituation effects. *International Journal of Social Robotics* 1, 2 (2009), 141–155.

[126] Carly Sitrin. 2016. Have Roombas become a part of the family? *The Boston Globe* (15 Sept. 2016). https://www.bostonglobe.com/lifestyle/style/2016/09/15/our-bots-ourselves/KekBWFnovSSp2yAhaTUOKN/story.html

[127] Vasant Srinivasan and Leila Takayama. 2016. Help me please: Robot politeness strategies for soliciting help from humans. In *Proceedings of the 2016 CHI conference on human factors in computing systems*. 4945–4955.

[128] Aaron Steinfeld, Terrence Fong, David Kaber, Michael Lewis, Jean Scholtz, Alan Schultz, and Michael Goodrich. 2006. Common metrics for human-robot interaction. In *Proceedings of the 1st ACM SIGCHI/SIGART conference on Human-robot interaction*. 33–40.

[129] Lucille Alice Suchman. 2007. *Human-machine reconfigurations: Plans and situated actions*. Cambridge university press.

[130] Leila Takayama, Doug Dooley, and Wendy Ju. 2011. Expressing thought: improving robot readability with animation principles. In *Proceedings of the 6th international conference on Human-robot interaction*. 69–76.

[131] Xiang Zhi Tan, Marynel Vázquez, Elizabeth J Carter, Cecilia G Morales, and Aaron Steinfeld. 2018. Inducing bystander interventions during robot abuse with social mechanisms. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. 169–177.

[132] Fumihide Tanaka, Aaron Cicourel, and Javier R Movellan. 2007. Socialization between toddlers and robots at an early childhood education center. *Proceedings of the National Academy of Sciences* 104, 46 (2007), 17954–17958.

[133] Ana Tanevska, Shruti Chandra, Giulia Barbareschi, Amy Eguchi, Zhao Han, Raj Korpan, Anastasia K Ostrowski, Giulia Perugia, Sindhu Ravindranath, Katie Seaborn, et al. 2023. Inclusive hri ii: equity and diversity in design, application, methods, and community. In *Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*. 956–958.

[134] Sam Thellman, Annika Silvervarg, Agneta Gulz, and Tom Ziemke. 2016. Physical vs. virtual agent embodiment and effects on social interaction. In *International conference on intelligent virtual agents*. Springer, 412–415.

[135] Rafael Thomas-Acaro and Brian Meneses-Claudio. 2024. Technological assistance in highly competitive sports for referee decision making: A systematic literature review. *Data and Metadata* 3 (2024), 188–188.

[136] Lucien Tisserand, Brooke Stephenson, Heike Baldauf-Quilliatre, Mathieu Lefort, and Frédéric Armetta. 2024. Unraveling the thread: understanding and addressing sequential failures in human-robot interaction. *Frontiers in Robotics and AI* 11 (2024), 1359782.

[137] Songül Tolan. 2019. Fair and unbiased algorithmic decision making: Current state and future challenges. *arXiv preprint arXiv:1901.04730* (2019).

[138] Jacqueline Urakami and Katie Seaborn. 2023. Nonverbal cues in human–robot interaction: A communication studies perspective. *ACM Transactions on Human-Robot Interaction* 12, 2 (2023), 1–21.

[139] Michael Veale and Reuben Binns. 2017. Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data. *Big Data & Society* 4, 2 (2017), 2053951717743530.

[140] Joshua Wainer, David J Feil-Seifer, Dylan A Shell, and Maja J Mataric. 2006. The role of physical embodiment in human-robot interaction. In *ROMAN 2006-The 15th IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, 117–122.

[141] Shengye Wang and Henrik I Christensen. 2018. Tritonbot: First lessons learned from deployment of a long-term autonomy tour guide robot. In *2018 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 158–165.

[142] Patrick P Weis and Cornelia Herbert. 2022. Do I still like myself? Human-robot collaboration entails emotional consequences. *Computers in Human Behavior* 127 (2022), 107060.

[143] Brenda K Wiederhold. 2024. Humanity's evolving conversations: AI as confidant, coach, and companion. *Cyberpsychology, Behavior, and Social Networking* 27, 11 (2024), 750–752.

[144] Jang Wonseok, Kang Young Woo, and Kang Yeonheung. 2021. Who made the decisions: Human or robot umpires? The effects of anthropomorphism on perceptions toward robot umpires. *Telematics and Informatics* 64 (2021), 101695.

[145] Sarah Woods, Kerstin Dautenhahn, Christina Kaouri, Renete Boekhorst, and Kheng Lee Koay. 2005. Is this robot like me? Links between human and robot personality traits. In *5th IEEE-RAS International Conference on Humanoid Robots*,

2005. IEEE, 375–380.

[146] Baoqin Wu, Muhammad Afzaal, and Dina Abdel Salam El-Dakhs. 2025. 'Yet his silence said volumes': a pragmatic analysis of conversational silence in rapport management. *Cogent Arts & Humanities* 12, 1 (2025), 2451490.

[147] Xiaolong Wu, Shuhua Li, Yonglin Guo, and Shujie Fang. 2024. Human or AI robot? Who is fairer on the service organizational frontline. *Journal of business research* 181 (2024), 114730.

[148] Ricarda Wullenkord and Friederike Eyssel. 2020. Societal and ethical issues in HRI. *Current Robotics Reports* 1, 3 (2020), 85–96.

[149] Claire Yang, Heer Patel, Max Kleiman-Weiner, and Maya Cakmak. 2025. Preserving Sense of Agency: User Preferences for Robot Autonomy and User Control across Household Tasks. *arXiv preprint arXiv:2506.19202* (2025).

[150] John Zerilli, Alistair Knott, James Maclaurin, and Colin Gavaghan. 2019. Transparency in algorithmic and human decision-making: is there a double standard? *Philosophy & Technology* 32, 4 (2019), 661–683.