CHALMERS
UNIVERSITY OF TECHNOLOGY

Responsible Humanoids: A Contradiction in Terms?

Citation for the original published paper (version of record):
Lemaignan, S., Moon, A., Coghlan, S. et al (2026). Responsible Humanoids: A Contradiction in Terms?. Hri 2026 Proceedings of the 21st ACM IEEE International Conference on Human Robot Interaction: 1341-1349. http://dx.doi.org/10.1145/3757279.3788817

Responsible Humanoids: A Contradiction in Terms?

Séverin Lemaignan
PAL Robotics
Barcelona Spain
IIIA-CSIC
Barcelona Spain
severin.lemaignan@iiia.csic.es

AJung Moon
McGill University
Montreal Canada
ajung.moon@mcgill.ca

Simon Coghlan
University of Melbourne
Melbourne Australia
simon.coghlan@unimelb.edu.au

Emily C. Collins
Univerisity of Manchester
Manchester United Kingdom
e.c.collins@manchester.ac.uk

Vanessa Evers
University of Twente
Enschede Netherlands
Centrum Wiskunde en Informatica
Amsterdam Netherlands
NTU
Singapore Singapore
vanessa.evers@ntu.edu.sg

Nico Hochgeschwender
University of Bremen
Bremen Germany
nico.hochgeschwender@uni.lu

Sara Ljungblad
University of Gothenburg
Gothenburg Sweden
Chalmers University of Technology
Gothenburg Sweden
sara.ljungblad@chalmers.se

Michael Milford
Queensland University of Technology
Brisbane Australia
michael.milford@qut.edu.au

Sarah Moth-Lund Christensen
University of Sheffield
Sheffield United Kingdom
s.m.l.christensen@sheffield.ac.uk

Francisco J. Rodríguez Lera
University of León
León Spain
fjrodl@unileon.es

Pericle Salvini
University of Oxford
Oxford United Kingdom
pericle.salvini@cs.ox.ac.uk

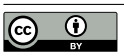
Yi Yang
KU Leuven
Leuven Belgium
yi.yang@kuleuven.be

Abstract

In this paper, we critically examine the current "humanoid hype" in robotics, questioning its alignment with responsible robotics principles. While technical challenges drive internal fascination, the pervasive public image of humanoids demands deeper HRI engagement. We explore how responsible robotics concepts, such as privacy, dignity, and trust, are uniquely challenged or overlooked in the pursuit of anthropomorphic robot forms. By dissecting this hype, and mapping the main findings of the recently-published *Roadmap for Responsible Robotics* to the humanoids field, we aim to move beyond technical form-factor obsessions to understand the true societal implications and identify potential blind spots for the HRI community.

CCS Concepts

• **Computer systems organization → Robotics**; • **Human-centered computing → HCI theory, concepts and models**.



HRI '26, Edinburgh, Scotland, UK

ACM ISBN 979-8-4007-2128-1/2026/03
https://doi.org/10.1145/3757279.3788817

Keywords

Responsible robotics, Humanoids

ACM Reference Format:
Séverin Lemaignan, AJung Moon, Simon Coghlan, Emily C. Collins, Vanessa Evers, Nico Hochgeschwender, Sara Ljungblad, Michael Milford, Sarah Moth-Lund Christensen, Francisco J. Rodríguez Lera, Pericle Salvini, and Yi Yang. 2026. Responsible Humanoids: A Contradiction in Terms?. 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, 9 pages. https://doi.org/10.1145/3757279.3788817

1 The Allure of Humanoids: Public Fascination vs. Pragmatic Reality

There is a temptation for roboticists and Human-Robot Interaction (HRI) practitioners to believe there should be more robots–including very human-like ones–in the society of tomorrow [21]. Riek and Irani call out five arguments underlying this position, including that "the jobs being automated are undesirable," that "labor shortages necessitate automation," and that "automation will give us all more leisure time" [31]. Investment in highly anthropomorphic robots or *humanoids* that piggyback on the global AI race, provides significant impulse to these arguments. Indeed, a growing number of voices now argue that we need to build and deploy humanoids, as opposed to robotic systems without human form

or characteristics–because: (a) a human-like form is most efficient for building robots that can perform undesirable or short-staffed human jobs in human work environments [5, 32]; (b) humanoids, unlike non-humanoid robots, represent general-purpose systems that can do everything that humans can do [18, 26]; and (c) humanoids are a kind of "physical-AI" that can unlock the next stage of the global AI race that has previously been limited to text and image data and generation [19, 27, 32]. As we outline further below, fueling this line of thinking is an impressive amount of both public and private investment.

In this paper, we question this logic. We analyze the hype over humanoids and propose how to view these anthropomorphic machines through the lens of 'responsible robotics', drawing on the recently published *Roadmap for Responsible Robotics* [7]. In doing so, we point to critical issues that need to be understood and accommodated in robotics and HRI [11]. Our paper is a call to action for HRI practitioners to be more conscious of the way we navigate the hype around humanoids; if we can deeply appreciate the importance of undertaking ethical evaluations at the outset and learn from its results, then our collective efforts in robotics might do more good than harm.

Much of the robotics community (and certainly the HRI community) still primarily understand the term *humanoid* as signifying a form factor of a robotic platform and the associated engineering challenges (e.g., locomotion, manipulation). In HRI, what makes a robot considered to be a humanoid is its degree of anthropomorphism, which is determined by a combination of four morphological factors–surface appearances (e.g., presence of humanlike skin or hair/eyelashes), body-manipulators (e.g., presence of torso, hands, fingers), facial features (e.g., eyes, mounth), and mechanical locomotion (e.g., wheels, treads) [30]. This is independent of the functionality that a robot may have or signify.

Humanoid robotics, a subfield of robotics, has been focused on the engineering challenge of matching the humanlike appearance of a robot with the functionality that it signals [37]. A representative conference in the subfield, the IEEE International Conference on Humanoid Robots, has been around since year 2000 [2]. For the last two decades, the subfield has always been a testbed for complex control algorithms, especially on locomotion and manipulation. Competitions such as RoboCup [3], a robot soccer competition involving bipedal systems, and DARPA Robotics Challenge [1] have historically supported and celebrated engineering advancements in the domain.

But in today's 2025 collective psyche, the word *humanoid* now embodies more: the promise of a robot that is general-purpose and generally useful. This semantic shift has been fueled by a combination of technical advancements, media portrayals, and business-driven storytelling.

1.1 Humanoids as the Original "Robots"

The fascination with humanoids in the Western culture traces back to the 1920s play *R.U.R.* or *Rossum's Universal Robots* [10] by Karel Čapek, a playwright known for societal satire. The play is often credited as the origin of the word, "robot." In the play, some businessmen create a factory that builds artificial beings (robots) out of chemical synthesis and electromechanical systems. The robots in the story are intentionally made to be highly anthropomorphic–in fact, there are no mentions of other types of robots such as robot dogs. The characters' hope in making robots mirrors some of the dominant arguments for robotics today–including that robots will perform work better than humans while relieving people of burdensome toil.

As Christoforou and Müller notes in their reflection of *R.U.R* [12], much of the societal issues depicted in the story applies to the modern society: Robot production is rather monopolized; Governments become interested in humanoids for military purposes; There are concerns about the sovereignty of the robots where each country seeks to have their own national (rather than universal) robot.[1] The story ends as a tragedy where the robot makers do not regret having made them, even as robots end their lives and the civilization. In a similar vein, the classic movie Metropolis (1927) depicts an evil robotic copy of a woman who fought for workers' rights in a dehumanizing system.

These are, of course, stories. However, given the eerie amount of commonality the 1920s play seems to have with the societal reality of today–over a hundred years later–, we agree with Christoforou and Müller that proactive approach to consider robot ethics is the most promising strategy for the society.

1.2 The Humanoid Hype

As some roboticists have publicly expressed, it is at least arguable that humanoids are not the ideal to solve human problems and needs (e.g., [9]). Yet, humanoids are one of the most active and heavily invested technology fields of today. In terms of trajectory, it is even outstripping the investments in emerging technologies such as autonomous vehicles[2] and even AI after the public launch of foundational AI models via ChatGPT in November 2022. Some estimate that humanoids represent a $5 Trillion USD market [26]. Governments have been aggressively supporting the private sector to be competitive in this market: In 2023, China's Ministry of Industry and Information Technology announced its ambitions to have two to three Chinese companies be global leaders of the industry by 2025 [6]; in 2025, the South Korean government announced its investment of 1 Trillion KRW (approx. $710M USD) toward making the country "the world's strongest country in the field of humanoid" [38]. The global race towards "physical-AI" is on [19].

Such mismatch between the over-enthusiasm or heightened expectation about a technology and the technology's maturity/readiness level is what characterizes a hype [16]. While the aforementioned commercial and pragmatic arguments for humanoids have existed in the past, several important factors have changed in the last few years that set the stage for today's humanoid hype.

Firstly, there have been genuinely remarkable technical advances. Legged motion is now technically much more advanced than one decade ago. Starting with quadruped robots–perhaps most notably Boston Dynamic's Big Dog, then Spot Mini, and now a plethora

[1]Note that South Korea's ambitious plans to become a leader in humanoids involve creating "K-humanoids," referring to the goal of establishing Korean-made humanoids [38].
[2]Approximately $100B USD have been spent directly or indirectly on autonomous vehicles domain [20].

of systems including from companies like Unitree, and graduating into bipedal humanoid robots–legged robots are now orders of magnitude better at walking, running, jumping. While manipulation and other functionalities needed to unlock interesting physical interactions in the real world are still far less advanced, many of us remember impactful demonstration videos posted by humanoid companies such as Boston Dynamics[3] and Unitree. Such media coverages have brought humanoid robots into the mainstream spotlight and fostered a renewed robotic imaginary.

Second, the cost and scale of humanoid manufacturing have changed radically. A humanoid might have cost millions of dollars a decade ago. Today, they are much more capable and affordable. At the time of writing, Unitree's humanoid, G1, is currently sold for $16K USD.[4] Such drop in price point makes humanoids accessible to wider market actors, tempting them to look for possible application areas. There are indeed some use cases where humanoids are especially sought. For instance, robots built for autistic children benefit from the humanlike form [33]. However, on the whole, only few use-cases for humanoid robots exist today, preventing the *grounding* of the place and identity of humanoid robots in concrete roles and tasks [9]. For instance, Agility Robotics stands out as one of the only companies with specific use-case for their humanoid robots (that of bin handling) [4].

Lastly–in addition to the advancements in design and manufacturing of powerful, electric-based actuation–there have been a rapid advancement in AI, including Reinforcement Learning (RL). AI-based control started to challenge the status quo of model-based control methods continues to fuel progress in humanoid whole-body control. The accelerated success of AI, rapid adoption of large language models, and the development of foundation models of robotics have set the stage for investors and governments alike to look to robotics as a next frontier [37].

Importantly, these achievements have been primarily driven by the private sector–mostly young companies–with a focus on building a business narrative that would support further investment (and, for many of these companies, ensure survival). This creates a unique situation where the public perception on one hand, and media portrayals and business narratives on the other hand, cross-feed to exacerbate the expectations on what humanoid robots are, and what they can do.

There is a sense that–much like the dynamics of AI research–research and development in humanoid robotics is shifting its weight from the laboratory to the industry, moving at the same time from the hands of academic roboticists to those of engineers and AI practitioners. As this transition takes its course through the patterns of a hype cycle, we argue that it is time for the robotics community to seriously consider what responsible robotics would mean in this context and how to navigate the hype as responsible roboticists.

[3]Boston Dynamics showed their Atlas robot doing back-flips in a video in 2017, https://www.youtube.com/watch?v=fRj34o4hN4I

[4]Unitree, https://shop.unitree.com/

1.3 Responsibility and Humanoids: A Complex Intersection

Responsible robotics refers to the notion that all the parties involved in any part of the life cycle of a robotic system–i.e., not only those involved in development, deployment, and integration but also in usage and maintenance–need to act in a responsible manner toward all stakeholders [7]. In this paper, we critically examine what responsible robotics looks like at a time of gold rush toward developing humanoids, whether *responsible humanoids* can even be conceptualized. To do so, we examine the humanoid hype using the findings of the recently-published *Roadmap for Responsible Robotics* [7].

In [7], the concept of responsibility is decomposed into multiple axes that constitute the foundation for future research and governance in the field. These include *trust*, ensuring that robots operate reliably and in alignment with legitimate human interests; *justice and fairness*, preventing inequitable or exploitative outcomes across contexts of design, deployment, and decommissioning; *dignity*, respecting the intrinsic worth, vulnerability, and autonomy of human beings in all interactions with robots; *environmental sustainability*, *privacy*, and *safety*, each reflecting a core ethical imperative for robotics as an embodied technology; and finally constructs commonly shared with Responsible AI frameworks [23]: *accountability*, securing traceable responsibility for decisions and actions throughout the robot's lifecycle, *transparency*, *understandability*, and *predictability*, further supporting the development of socially intelligible robotic systems. Collectively, these dimensions delineate an integrative ethical framework through which robotics may co-evolve with human societies in a manner that safeguards human agency and upholds humane values.

In the view of the roadmap's authors, responsible robotics extends but also departs from "responsible AI" by emphasizing the distinct socio-technical challenges associated with robots' embodiment, autonomy, and physical interaction with the world [25]. The concept requires that stakeholders – including researchers, engineers, policymakers, industry representatives, and end-users – act conscientiously throughout the robot lifecycle, ensuring that robotic technologies promote human well-being, social justice, and ecological integrity rather than undermining them.

These axes are not new to the HRI community, which has long engaged with them in various forms. However, the current humanoid hype presents a unique case where these principles are both critically challenged and often overlooked, due to the combination of technical fascination and socio-economical incentives that we have described above.

2 Mapping Responsible Robotics Principles to Humanoid Development

In this section, we map key concepts identified in the Responsible Robotics roadmap to the specific challenges and considerations that arise in the context of humanoid robots.

2.1 Trust Without Blindness: Fostering Calibrated Interactions

The Roadmap for Responsible Robotics [7] warns that people may consciously or subconsciously overtrust robotic systems. It calls for a realistic and nuanced understanding of trust as it relates to the design, adoption, and use of robots, recognizing its close ties to reliability, understandability, justice, and privacy. The Roadmap also emphasizes that trust is inherently contextual. Different stakeholders, such as developers, operators, and end users, evaluate the same robot according to different expectations. Responsible robotics therefore aims not to maximize trust but to promote calibrated trust that aligns perception with demonstrated performance.

Calibrated trust is particularly difficult to achieve in humanoid robots. Their anthropomorphic appearance and expressive behavior often elicit trust before reliability is proven. A humanoid's gaze, tone, or gesture activates the same social heuristics humans use to interpret intention and emotion. When these cues are simulated without genuine understanding, trust shifts from an evidence-based judgment to an automatic emotional response. The same design elements that make humanoids appear trustworthy can thus foster overreliance, emotional dependency, and even manipulation. In healthcare or companionship contexts, a humanoid that appears to care may encourage uncritical attachment, while its actual objectives and data flows are controlled by distant manufacturers or service providers.

This raises a fundamental question about the object of trust, especially when humanoids are framed as potential teammates in human–robot teams. Effective collaboration requires that humans understand what the robot will do, rely on it, and delegate responsibilities safely. They need a teammate that is transparent, accountable, and reliable. However, when that "teammate" is a humanoid, this desire becomes a vulnerability. The humanoid's form creates a powerful illusion of shared intent, where humans engage as if in genuine cooperation while the robot merely executes an optimized program. Even formal modeling techniques such as Brahms notation [35] can capture an idealized team workflow but fail to resolve the fundamental asymmetry when one agent only simulates partnership. When trust is distributed across layers of embodiment, code, and corporate infrastructure, it ceases to be a stable property of the robot and becomes a moving target embedded in opaque accountability networks. Under these conditions, the ethical call to build trustworthy systems risks becoming a marketing exercise that produces trust-inducing rather than trustworthy machines.

Therefore, responsible humanoid design should move from eliciting trust to enabling it. Designers should not simulate confidence but instead make the robot's limitations, uncertainties, and control boundaries explicit through transparent behavior, interpretable feedback, and clear contextual communication. Trust should emerge from verifiable performance and openness rather than from human-like appearance or behavior. Trust is not a by-product of anthropomorphism but a result of epistemic honesty. Responsible robotics should aim not to make users believe in robots but to ensure that they know when and why a robot deserves to be trusted.

2.2 Justice and Fairness: Addressing Access, Displacement, and Exploitation

The Roadmap for Responsible Robotics [7] emphasizes that robots, like other technologies, raise significant issues of justice and fairness throughout their entire lifecycle. These concerns extend beyond algorithmic bias and data ethics to include the material and social dimensions of robotics. Fairness must therefore be addressed from design and production through deployment, maintenance, and eventual decommissioning. The Roadmap highlights several key questions. Are the materials and components used in robotic systems ethically sourced, or do they rely on extractive and exploitative supply chains? Are labor practices within robotics companies fair, and do they distribute risks and benefits equitably? During deployment, do systems privilege certain groups while excluding others due to cost, accessibility, or cultural assumptions? Responsible robotics, in this sense, requires continuous reflection and adaptation rather than a one-time ethical checklist.

In the case of humanoid robots, these justice and fairness considerations acquire particular urgency. Humanoids are often developed and manufactured in wealthy, industrialized contexts but are imagined as future workers or companions in diverse global environments. Their design frequently reflects Western cultural norms of embodiment, gender, and labor, which may not translate fairly or respectfully to other settings. As a result, the global pursuit of humanoids risks reproducing historical inequalities in which technological benefits concentrate in the Global North while material and social costs are externalized elsewhere. For example, the rare-earth minerals required for high-performance actuators and sensors are often sourced from regions facing exploitative mining conditions. Similarly, the high financial cost of humanoids limits their availability to affluent institutions, thereby reinforcing socioeconomic disparities in who benefits from automation.

Justice and fairness also extend to questions of labor and displacement. Promoters of humanoids often justify their development as a solution to undesirable or low-status jobs. However, this framing obscures the human consequences of job loss and the systemic structures that define such work as undesirable in the first place. A humanoid robot that replaces a care worker or cleaner may alleviate a labor shortage, but it also simultaneously devalues care as a human vocation and erases the livelihoods of those performing it. This dynamic perpetuates a cycle in which automation promises fairness through efficiency while reproducing injustice through exclusion and dispossession.

Ensuring justice in humanoid robotics requires attention to both global and local contexts. Designers and policymakers must critically examine who gains and who loses at each stage of a humanoid's lifecycle. Fairness should be operationalized not only through compliance or inclusion metrics but through active participation of affected communities in design and policy processes. As the Roadmap makes clear, justice is not achieved by a single ethical decision during development. It is sustained through ongoing reflection, equitable distribution of benefits and harms, and continuous adaptation as humanoid technologies evolve.

2.3 Accountability in the Age of Humanoids: Bridging Regulatory Gaps

While regulatory initiatives such as the EU AI Act or the U.S. AI Action Plan define external frameworks for accountability, a significant portion of the resistance to operationalizing responsible humanoid robotics arises from internal drivers within the research and engineering communities. These include persistent technical challenges, such as limited scalability of embodied intelligence models, fragile locomotion control, and insufficient interoperability across robotic architectures. Moreover, research inertia – the tendency to reproduce established benchmarks, architectures, and experimental paradigms – slows the integration of new practices that embed transparency, traceability, and explainability from the design phase. Consequently, even well-intentioned research teams often struggle to align innovation speed with compliance requirements and also with individual needs and HRI contexts.

This gap between regulatory ambition and technical maturity is further reflected in the current software landscape of humanoid robotics. Although several humanoid robots have recently entered pre-industrial or large-scale production, their software architectures remain largely prototypical. In most cases, the hardware is being standardized faster than the underlying control and interaction software, which still follows a proof-of-concept logic rather than a defined Software Development Lifecycle (SDLC). Long-term maintenance policies, dependency management, and security update mechanisms are seldom documented, and the same codebases are frequently repurposed across radically different contexts – from industrial environments to eldercare facilities or educational spaces – without systematic adaptation or validation. This mismatch between manufacturing maturity and software immaturity exposes a critical gap in accountability and lifecycle governance, challenging the applicability of emerging regulatory frameworks such as the AI Act.

To ensure better accountability, robotics community must align innovation practices with emerging regulatory and ethical standards, and actively engage in return with these bodies. In this sense, the current phase of humanoid expansion represents both a technical challenge and a social experiment in how responsibility, compliance, and ambition can co-evolve within the global robotics ecosystem.

2.4 Sustainability Forgotten? The Lifecycle of Humanoid Robotics

The Roadmap for Responsible Robotics [7] urges the robotics community to consider sustainability across a robot's entire lifecycle, from material sourcing and energy use to repair and end-of-life planning. While these concerns apply broadly, the current humanoid boom prioritizes performance and human-likeness over ecological responsibility. The humanoid form may be one of the least sustainable configurations in robotics, favoring aspiration and marketing over efficiency and material restraint.

Energy efficiency is a central challenge. Bipedal locomotion demands constant control and high power simply to stay upright, making it far less efficient than wheeled or fixed-base systems. These costs are compounded by data-driven control models trained and run on energy-intensive computing infrastructure. Humanoids reveal a clear tension between autonomy and sustainability: their most advanced capabilities often depend on computational and material resources that far exceed their social or practical benefit.

Humanoids also embody the material and manufacturing burdens of high-performance design. Their use of specialized alloys, rare-earth elements, and carbon composites links them directly to extractive and energy-intensive global supply chains. Rapid iteration and competition foster short-lived prototypes with little potential for reuse. When obsolete, these machines often become expensive electronic waste, difficult to repair or recycle because of their tightly integrated construction.

Repairability and reuse remain the weakest points in the humanoid lifecycle. Most platforms are proprietary and closed, limiting access to maintenance tools and replacement parts. Even minor failures can require costly manufacturer intervention, discouraging repair and accelerating obsolescence. The result is a paradox: machines designed to emulate human resilience but engineered for short operational lives.

Humanoids, therefore, pose a critical test for responsible robotics. Can a technology modeled on the human body also embody environmental care? Meeting the principles of the Roadmap requires more than symbolic gestures toward "green robotics". It calls for measurable reductions in energy use, transparent supply chains, and truly open and repairable architectures. Without these commitments, humanoid robotics risks becoming a case study in how technological ambition outpaces ecological accountability.

2.5 Privacy in Motion: Humanoids as Perpetual Surveillance

The Roadmap for Responsible Robotics [7] warns that robots that are equipped with perception mechanisms and networked data systems can easily become active trackers. While privacy-by-design principles call for proactive, user-centered protection of data, the embodied nature of robots introduces new challenges. Unlike disembodied AI systems, humanoid robots operate within shared human spaces, collecting and processing information not only about their users but also about bystanders who have not consented to being observed. In such contexts, privacy violations occur not only through stored data but through the robot's very presence and gaze.

Humanoid robots transform surveillance into interaction. Their cameras and microphones are embedded in familiar anthropomorphic forms. When a humanoid nods, follows movement, or makes eye contact, these gestures function both as social engagement and as data collection. What seems like a relational encounter is also a sensor event that may be processed, stored, and transmitted beyond the immediate environment.

In such interactions, observation and communication are inseparable, undermining traditional ideas of informed consent. Unlike digital systems where surveillance occurs in the background, humanoid monitoring happens in the foreground: the act of speaking, moving, or simply being present produces data. Users cannot meaningfully choose whether to be observed when monitoring is built into the interaction itself. This creates a consent illusion that voluntary engagement equals informed agreement.

Beyond data governance, humanoid presence reshapes the spatial and social dynamics of privacy. Robots deployed in homes,

hospitals, classrooms, or workplaces extend institutional observation into spaces once considered private, often under the guise of care, safety, or assistance. Their mobility and autonomy collapse traditional boundaries between public and private, creating a state of continuous perceptual exposure. The sense of being constantly within a robot's field of view introduces a new dimension of privacy loss, which is difficult to measure.

Hence, responsible humanoid design must expand the concept of privacy beyond compliance and encryption. It must account for how embodiment, perception, and proximity alter human experiences of being seen. Designing for privacy means making sensing visible, making consent revocable, and making presence negotiable. A responsible humanoid makes its sensing transparent, allowing people to know and control when they are observed.

2.6 Safety in Close Quarters: Navigating Physical and Psychological Risks

The safety and reliability of humanoid robotics must be addressed at both the embodiment and overall system levels. Humanoid robots impose a new class of dominant risks, such as falls and loss of balance, which calls for machine-specific safety standards for continuous balance control and coordinated motion relevant for complex physical robot-environment interactions that are currently not covered or sufficiently covered by existing mobile or industrial robot standards. Although humanoid embodiment contributes to novel safety risks, it is not sufficient to make the embodiment itself safe. A holistic safety perspective of the complete humanoid robot application, including the environment, task, and interaction context, is required, specifically as new humanoid applications and thus evolving risks (e.g., in domestic or industrial settings) are in reach. This follows the same shift already seen in collaborative robotics, where the focus has moved from cobot safety to application safety.

Another source of safety risks arises from the perceived agency and frequent overestimation of humanoid capabilities. When people attribute autonomy or understanding to a humanoid, they may over-trust it, delay intervention during malfunction, or misinterpret social cues as competence. Such reliance can reduce vigilance and lead to unsafe proximity, erroneous delegation, or confusion about responsibility in tasks. The result is a new category of safety risks, emerging not from hardware but from human perception and interaction.

Additional risks emerge from the employed approaches to robot control. Modern humanoid control increasingly depends on large, data-driven vision-language-action (VLA) models with impressive zero-shot generalization capabilities in novel task settings. However, the integration of these models into humanoid robots raises profound safety concerns. A frequently held assumption is that if these models are trained on safe data, their behavior will be safe. Although this may hold in restricted scenarios, it does not provide guarantees in dynamic and open-ended environments, where unexpected disturbances or out-of-distribution conditions are unavoidable. Simply scaling the amount of training data reduces but does not eliminate the risk of unsafe actions, and no formal guarantees can be derived from this approach. Responsible humanoid robotics therefore requires explicit safety scaffolds around learning-based controllers (e.g., runtime verification, safety guards, etc.) combined with clear communication that maintains accurate user expectations. Only by integrating application-level risk assessment, standardized safety procedures, and awareness of agency-induced social risks can humanoid robotics become both reliable and safe.

The expansion of robots' capabilities to interact with humans not only at the physical level but also at the cognitive, emotional, and social levels introduces new types of risks. These risks extend beyond mere physical harm (e.g., cutting, crushing, etc.) to include psychological and ethical harms, such as addiction, social isolation, abstraction, moral disengagement, and the decline of interpersonal skills [8]. Such risks may be further amplified by the humanoid appearance and lifelike presence of robots, which foster anthropomorphism and strengthen the illusion of emotional and social reciprocity between humans and machines.

2.7 Seeing Through the Mask: The Need for Transparency and Predictability

The increasing visibility of humanoid robots in media and research demonstrations has created a paradox between appearance and capability. While their external design evokes human-level competence, their actual cognitive and functional abilities remain narrow, brittle, and context-dependent. This paradox does not only shape public perception—it also permeates the robotics research community itself, particularly within Human–Robot Interaction. The field often oscillates between scientific rigor and performative display: carefully staged experiments and conference demonstrations may prioritize engagement, novelty, or funding visibility over reproducibility and technical transparency. In an environment driven by competitive project calls, industrial partnerships, and media exposure, even well-intentioned researchers may inadvertently contribute to the myth of imminent general-purpose humanoids. Recognizing this internal amplification of expectations is essential for establishing communication standards that balance scientific credibility with public engagement. To foster trustworthy interaction, humanoid systems must therefore become not only technically reliable but also transparent and predictable – capable of conveying what they can and cannot do, and under what conditions.

To support responsible design and deployment of humanoids, it is essential to establish transparent communication about their capabilities and limitations. Effective and safe interaction depends not only on technical reliability but also on users' informed understanding of what a system can and cannot do. Structured approaches to communication may serve to operationalize this transparency by distinguishing between the content of information and the manner of its delivery. Recent work complements the IEEE 7001-2021 standard by classifying transparency into Content and Interaction dimensions, the latter encompassing task, natural interaction, and system intention aspects [28]. However, transparency alone is insufficient without adequate user training, which ensures that operators and end-users can accurately interpret system behavior, recognize operational boundaries, and respond appropriately to errors or ambiguities. Integrating training programs alongside transparent design practices is therefore essential to cultivate realistic user expectations and promote responsible engagement with

humanoid technologies, especially in high-stake domains such as healthcare and education [24].

2.8 Dignity Under Pressure: Ethical Red Lines for Humanoid Applications

In the Responsible Robotics Roadmap, dignity is defined as the ethical imperative to respect the intrinsic worth, vulnerability, and autonomy of human beings in all human–robot interactions. The authors of the Roadmap emphasise that robotic systems must preserve individuals' sense of agency and social recognition, and that robots (e.g., in warfare or the sex industry) that commodify, degrade, or devalue human life should be avoided. Dignity thus serves as a moral boundary for acceptable robotic applications, an ethical red line, linking technological design to broader commitments to peace, respect, and human rights.

When near-future humanoid robots are pitched to the general public, it is often as a dependable caregiver for elderly relatives, as a home-maker doing all the boring chores, or as a compliant sexual partner. Due to their physical appearance and other resemblances, it is perhaps unsurprising that humanoid robots would be considered for placement in human environments in lieu of actual human beings [36]. This placement could involve primarily non-public environments such as private homes or care homes [39]. It could also involve robots undertaking tasks that put users in unwanted or vulnerable situations. For example, a humanoid may be designed for and inserted into a sensitive environment in which the robot imposes on the user's space and well-being, either physically or through its data collection. It may be tempting to believe that, for example, a lonely and socially isolated older person will benefit from being given a humanoid robot companion. However, it is evident that some older people who live alone consider the imposition of certain robot companions as not helpful, and, moreover, as infantalising and undignified [13]. While not all people may feel this way, the assumption that humanoid companions will always benefit those needing care or emotional connection must be avoided. As with all technology a degree of personal choice and flexibility of use should be considered, with fear of exploitation and user preference, balanced against the benefits technology can bring to independent living [14].

By their very form and their AI-based abilities, humanoid robots can also invite a misguided kind of anthropomorphism [15]. For example, some people may project human characteristics onto a robot such as advanced agency, personhood, or general intelligence, which the robot lacks. This, in conjunction with the earlier mentioned trust and transparency concerns, may raise ethical concerns about rendering users emotionally, socially, and epistemically vulnerable. Such humanoids could make users susceptible to unsafe interactions or manipulation by unscrupulous actors.

In the development and deployment of humanoid robots, reflection on how to preserve human dignity in robot-human interaction is paramount [34]. It forces us to ask fundamental questions in earlier ideation and development phases about the appropriateness of the humanoid form for the later deployment context. Here we can draw upon lessons learned in the Responsible AI literature [17] and champion co-design principles [29], thereby actively involving potential user feedback and reflections. Or relatedly, integrate humanities and arts methodologies throughout the robot life cycle [22], thereby foregrounding commitments to ongoing reflection and ethical principles such as dignity and autonomy.

3 Conclusion: Towards a Critical and Proactive HRI Engagement

3.1 The Absence of HRI in Mainstream Humanoid Discourse

Humanoids, like autonomous vehicles over the past decade, and more recently generative AI and so-called foundation models, are at peak tech hype and investment interest – with funding, talent and interest following accordingly. One of the challenges, as we have seen in these past tech waves (some of which are admittedly still ongoing), is that the approach, largely driven by commercial appetite, has not necessarily been an all-rounded one. The big players in humanoids – primarily companies and startups from China and the United States, bring different approaches to developing and promoting their humanoid technology development, but one thing in common is that rigorous, substantive considerations of factors like Human Robot Interaction have been largely absent to date. Instead, typical demos have heavily focused on robots mock-fighting humans in boxing or martial arts displays, or robots operating in manufacturing or warehouse logistics type environments. There are many current and potential issues with this lack of deep HRI consideration. One is simply that the technology stack may get too far without sufficient consideration of HRI factors, which will then have to awkwardly, and probably not effectively, be re-worked back into these systems. Another is around human expectations on the technology – without a proper consideration of HRI from an early stage, incorrect or inappropriate expectations may be baked into the public pscyche as pertains to these systems, which will not be helpful for any of the stakeholders in the long run.

Despite the clear relevance of human–robot interaction (HRI) to humanoid systems, these discussions remain largely absent from the core discourse in the humanoids research and development community. Current conversations focus mainly on mechanical design, locomotion, and control architectures, rather than on how humans will perceive, trust, or collaborate with these robots. This narrow focus risks producing humanoids that are technically impressive but socially and ergonomically unfit for real-world environments. Without early and systematic integration of HRI perspectives, developers may repeat past mistakes seen in autonomous vehicles and generative AI – advancing performance while neglecting usability, ethics, and societal alignment.

3.2 Operationalizing Responsible Robotics for Humanoids

The Responsible Robotics Roadmap outlines concrete gaps and opportunities for operationalizing responsible robotics in practice. Table 1 highlight those identified as priority areas for the humanoid robotics community, also putting in perspective the relevant stakeholders – illustrating the interdisciplinary nature of the challenges.

Responsible Robotics Gap	Stakeholders	Humanoid-Specific Challenge	Horizon
1. Identifying progress indicators	Researchers (social sciences), Standards committees with industry input	Rapid pace of humanoid development makes it difficult to define timely metrics reflecting both technical advances and societal impact.	short-term
2. Requirements Engineering	Interdisciplinary researchers, Software engineers	Translating societal expectations and ethical principles into concrete technical requirements is challenging due to multi-faceted socio-economical expectations built around humanoid robots.	short-term
3. Teaching Responsible Robotics	Educators, Researchers	Ensuring curricula capture both cutting-edge technical skills and broader societal, ethical, and legal considerations despite rapid field evolution.	short-term
4. Operationalizing explainability, predictability, understandability	Engineers, Philosophers	Humanoids' complex human interactions make it difficult to design behaviors that are easily interpretable without limiting functionality.	short-term
5. Reporting irresponsible incidents/practices	Legislators, Regulators, Professional associations	Public perception of humanoids as autonomous agents creates pressure for clear reporting systems, yet technical nuance is often lost in societal discourse.	medium-term
6. Responsible technological intervention in systemic problems	Policy experts, Engineers, Social Scientists, Philosophers, Citizens	Balancing resource constraints and potential for humanoids to address societal challenges requires careful prioritization and ethical oversight.	medium-term
7. Planning with interacting values in uncertain environments	Researchers (AI, Computer Science)	Humanoids must navigate conflicting goals and ethical norms in real time, a technically and socially complex problem.	medium-term
8. Testbeds for assessing interaction-based/ethical harms	Researchers (HRI, Psychology, Ethics), Users	Designing realistic and safe environments to study harmful interactions is difficult, especially when human expectations of humanoids vary widely.	medium-term
9. Enabling a second-hand humanoid market	Insurance, Regulators, Business, Roboticists	Liability, safety, and maintenance issues are heightened due to humanoids' autonomous and physical capabilities.	medium-term
10. Educating users to live with humanoids	Researchers (HRI, Education, Psychology)	Users often overestimate capabilities or misjudge risks, creating a need for structured education and guidance.	medium-term
11. Creation and curation of ML training datasets	Engineers, Policy experts, Philosophers, Citizens	Data must reflect ethical and societal norms, yet technical constraints and proprietary interests complicate consensus-building.	medium-term
12. Understanding and arbitrating trade-offs between goals and values	Policy experts, Regulators, Engineers, Social Scientists, Philosophers, Citizens	Humanoids' multifunctional and human-like nature intensifies ethical, social, and economic trade-offs, requiring ongoing negotiation between stakeholders and adaptive regulation.	long-term

Table 1: Responsible Robotics Gaps, Stakeholders, and Humanoid-Specific Challenges

3.3 Reclaiming the Narrative

At some point – without explicit consensus – the field of robotics seems to have collectively assumed that the humanoid form represents the inevitable endpoint of embodied AI. This silent convergence, reinforced by media narratives, industrial ambition, and funding incentives, has positioned humanoids as the default vision of our technological future. Yet this trajectory was never formally decided, debated, or ethically justified. The question "Responsible Humanoids: A Contradicton In Terms?" therefore demands not only technical reflection but also disciplinary self-awareness.

In reflecting on the trajectory of humanoid robotics, we must ask whether the field's collective imagination has already defaulted to a single morphological archetype – the human form. This implicit convergence may be more cultural than functional, driven by familiarity and anthropocentric bias rather than necessity. A responsible future for robotics should preserve morphological diversity, ensuring that human-centered design does not become synonymous with human-shaped design. Responsibility, in this sense, is not a property of the humanoid itself but a quality of the ecosystem that defines, builds, and deploys it. Collectively, the HRI community has a unique opportunity – and perhaps a responsibility – to reclaim, or at least, co-design, the narrative around humanoid robotics to place 'responsibility' at its core.

Position statement

We acknowledge that the current group of authors predominantly represents perspectives from the Global North and WEIRD (Western, Educated, Industrialized, Rich, and Democratic) research contexts. We highlight the urgent need to incorporate a broader range of viewpoints and greater diversity in future iterations or reformulations of similar critique. Our goal is to spotlight and examplify some critical issues, foster the advancement of responsible robotics, and facilitate our collective journey through the intricate socio-technical landscape ahead.

Acknowledgments

This work was made possible through Dagstuhl Seminar on *Roadmap for Responsible Robotics* (23371). In addition, we acknowledge the following sources of support for individual authors: Fisher is supported by the UK Royal Academy of Engineering's *Chair in Emerging Technologies* scheme and the work is partially funded by EPSRC in the UK, through the Computational Agent Responsibility project (EP/W01081X/1). Rodríguez-Lera by Grant PID2021-126592OB-C21 funded by MICIU/AEI/10.13039/501100011033 and by ERDF, EU; Ljungblad by Wallenberg AI, Autonomous Systems and Software Program, Humanity and Society; Moon by the Natural Sciences and Engineering Research Council of Canada; Hochgeschwender supported by the German Federal Ministry of Research, Technology and Space (BMFTR) under the Robotics Institute Germany (RIG).

References

[1] [n. d.]. *DARPA Robotics Challenge (DRC).* https://www.darpa.mil/research/programs/darpa-robotics-challenge
[2] [n. d.]. *Humanoid Robotics.* https://www.humanoid-robotics.org/
[3] [n. d.]. *RoboCup Humanoid League.* https://www.robocup.org/leagues/3
[4] Evan Ackerman. [n. d.]. Agility Robotics' Digit Is Getting Back to Work - IEEE Spectrum. ([n. d.]). https://spectrum.ieee.org/digit-agility-robotics
[5] Brett Adcock. 2022. *Roadmap to a Positive Future Powered by AI.* https://www.figure.ai/master-plan?__readwiseLocation=
[6] Digital Policy Alert. 2025. *China: Issued Guiding Opinions on the Innovation and Development of Humanoid Robots.* https://digitalpolicyalert.org/event/15378-issued-guiding-opinions-on-the-innovation-and-development-of-humanoid-robots Section: it.
[7] Dejanira Araiza-Illan, Kevin Baum, Helen Beebee, Raja Chatila, Sarah Moth-Lund Christensen, Simon Coghlan, Emily Collins, Kate Conroy, Alcino Cunha, Anna Dobrosovestnova, Hein Duijf, Vanessa Evers, Michael Fisher, Nico

Hochgeschwender, Nadin Kökciyan, Séverin Lemaignan, Francisco Rodriguez-Lera, Sara Ljungblad, Martin Magnusson, Masoumeh Mansouri, Michael Milford, AJung Moon, Thomas M Powers, Pericle Salvini, Teresa Scantamburlo, Nick Schuster, Marija Slavkovik, Ufuk Topcu, Daniel Vanegas, Andrzej Wasowski, and Yi Yang. 2025. A Roadmap for Responsible Robotics. *IEEE Robotics and Automation Magazine* (2025).

[8] British Standards Institution (BSI). 2023. *Robots and robotic devices. Ethical design and application of robots and robotic systems – Guide*. Technical Report / Draft Standard Project BS 8611 / Project 9021-05777. British Standards Institution. https://standardsdevelopment.bsigroup.com/projects/9021-05777 Published as BSI Standard BS 8611:2023; original project reference 9021-05777.

[9] Rodney Brooks. 2025. *Why Today's Humanoids Won't Learn Dexterity – Rodney Brooks*. https://rodneybrooks.com/why-todays-humanoids-wont-learn-dexterity/

[10] Karel Capek. 1973. *R.U.R.: Rossum's Universal Robots*. Pocket.

[11] Raja Chatila. 2019. *Inclusion of Humanoid Robots in Human Society: Ethical Issues*. Springer Netherlands, Dordrecht, 2665–2674. doi:10.1007/978-94-007-6046-2_147

[12] Eftychios G. Christoforou and Andreas Müller. 2016. R.U.R. Revisited: Perspectives and Reflections on Modern Robotics. 8, 2 (2016), 237–246. doi:10.1007/s12369-015-0327-6

[13] Simon Coghlan, Jenny Waycott, Amanda Lazar, and Barbara Barbosa Neves. 2021. Dignity, autonomy, and style of company: dimensions older adults consider for robot companions. *Proceedings of the ACM on human-computer interaction* 5, CSCW1 (2021), 1–25.

[14] Emily C Collins. 2017. Vulnerable users: deceptive robotics. *Connection Science* 29, 3 (2017), 223–229.

[15] Malene Flensborg Damholdt, Oliver Santiago Quick, Johanna Seibt, Christina Vestergaard, and Mads Hansen. 2023. A scoping review of HRI Research on 'anthropomorphism': contributions to the method debate in HRI. *International Journal of Social Robotics* 15, 7 (2023), 1203–1226.

[16] Ozgur Dedehayir and Martin Steinert. 2016. The Hype Cycle Model: A Review and Future Directions. 108 (2016), 28–41. doi:10.1016/j.techfore.2016.04.005

[17] Virginia Dignum. 2019. *Responsible artificial intelligence: how to develop and use AI in a responsible way*. Vol. 2156. Springer.

[18] Lauren Edmonds and Lakshmi Varanasi. 2025. *The story of Optimus, the humanoid robot at the heart of Elon Musk's growth plans for Tesla*. https://www.businessinsider.com/optimus-tesla-humanoid-robot-elon-musk-growth-plans-2025-9

[19] Fabrice R. Noreils. [n. d.]. *Humanoid Robots at Work: Where Are We ?* https://arxiv.org/html/2404.04249v1

[20] Jeff Farrah. 2025. *Autonomous Vehicles: Driving the Next Wave of Economic Growth*. https://www.theavindustry.org/blog/autonomous-vehicles-driving-the-next-wave-of-economic-growth Section: it.

[21] Mafalda Gamboa. 2025. Robots are Increasingly: Imagination Crisis in Human-Computer Interaction Research. In *Proceedings of the Sixth Decennial Aarhus Conference: Computing X Crisis (AAR '25)*. Association for Computing Machinery, New York, NY, USA, 216–222. doi:10.1145/3744169.3744189

[22] Drew Hemment, Cody Kommers, and et al. 2025. *Doing AI Differently: Rethinking the Foundations of AI via the Humanities*. Technical Report. London: The Alan Turing Institute.

[23] Anna Jobin, Marcello Ienca, and Effy Vayena. 2019. The Global Landscape of AI Ethics Guidelines. 1, 9 (2019), 389–399. doi:10.1038/s42256-019-0088-2

[24] Deepti Mishra, Karen Parish, Ricardo Gregorio Lugo, and Hao Wang. 2021. A framework for using humanoid robots in the school learning environment. *Electronics* 10, 6 (2021), 756.

[25] AJung Moon, Shalaleh Rismani, and HF Machiel Van der Loos. 2021. Ethics of Corporeal, Co-Present Robots as Agents of Influence: A Review. 2 (2021), 223–229.

[26] Morgan Stanley. 2025. *Humanoids: a $5 Trillion Market*. https://www.morganstanley.com/insights/articles/humanoid-robot-market-5-trillion-by-2050

[27] Kim Na-young. 2025. *S. Korea aims to enter mass production of humanoid robots in 2029, self-driving cars in 2030*. https://en.yna.co.kr/view/AEN20250910002700320 Section: Economy Business.

[28] Lejla Nukovic, Jérôme Kirchhoff, and Oskar von Stryk. 2024. Transparency Classification for HRI with Humanoid Service Robots. In *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction* (Boulder, CO, USA) *(HRI '24)*. Association for Computing Machinery, New York, NY, USA, 798–802. doi:10.1145/3610978.3640739

[29] Anastasia K Ostrowski, Cynthia Breazeal, and Hae Won Park. 2021. Long-term co-design guidelines: empowering older adults as co-designers of social robots. In *2021 30th IEEE international conference on robot & human interactive communication (RO-MAN)*. IEEE, 1165–1172.

[30] Elizabeth Phillips, Xuan Zhao, Daniel Ullman, and Bertram F. Malle. 2018. What Is Human-like?: Decomposing Robots' Human-like Appearance Using the Anthropomorphic roBOT (ABOT) Database. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction* (Chicago IL USA, 2018-02-26). ACM, 105–113. doi:10.1145/3171221.3171268

[31] Laurel D. Riek and Lilly Irani. 2025. The Future Is Rosie?: Disempowering Arguments About Automation and What to Do About It. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (Yokohama Japan, 2025-04-26). ACM, 1–14. doi:10.1145/3706598.3714151

[32] Arnaud Robert. 2025. *Embracing the Autonomous Future: How Humanoids Will Transform Industry - Hexagon Robotics*. https://robotics.hexagon.com/embracing-the-autonomous-future-how-humanoids-will-transform-industry/ Section: Articles.

[33] Bob R. Schadenberg, Dennis Reidsma, Vanessa Evers, Daniel P. Davison, Jamy J. Li, Dirk K. J. Heylen, Carlos Neves, Paulo Alvito, Jie Shen, Maja Pantić, Björn W. Schuller, Nicholas Cummins, Vlad Olaru, Cristian Sminchisescu, Snežana Babović Dimitrijević, Sunčica Petrović, Aurélie Baranger, Alria Williams, Alyssa M. Alcorn, and Elizabeth Pellicano. 2021. Predictable Robots for Autistic Children—Variance in Robot Behaviour, Idiosyncrasies in Autistic Children's Characteristics, and Child–Robot Engagement. *ACM Trans. Comput.-Hum. Interact.* 28, 5, Article 36 (Aug. 2021), 42 pages. doi:10.1145/3468849

[34] Amanda Sharkey. 2014. Robots and human dignity: a consideration of the effects of robot care on the dignity of older people. *Ethics and Information Technology* 16, 1 (2014), 63–75.

[35] Maarten Sierhuis, William J. Clancey, and Ron J.J. Van Hoof. 2007. Brahms: a Multi-agent Modelling Environment for Simulating Work Processes and Practices. *International Journal of Simulation and Process Modelling* 3, 3 (2007), 134–152. doi:10.1504/IJSPM.2007.015238

[36] Robert Sparrow. 2016. Robots in aged care: a dystopian future? *AI & society* 31, 4 (2016), 445–454.

[37] Yuchuang Tong, Haotian Liu, and Zhengtao Zhang. 2024. Advancements in Humanoid Robots: A Comprehensive Review and Future Prospects. 11, 2 (2024), 301–328. doi:10.1109/JAS.2023.124140

[38] Junho Yu. 2025. *The public and private sectors will invest more than 1 trillion won in the development of "K-humanoi.. - MK*. https://www.mk.co.kr/en/it/11288460 Section: it.

[39] Shuai Yuan, Simon Coghlan, Reeva Lederman, and Jenny Waycott. 2022. Social robots in aged care: Care staff experiences and perspectives on robot benefits and challenges. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (2022), 1–23.

Received 2025-10-17; accepted 2025-11-21