Towards Inclusive Robot Competitions

Zoe Evans, Muhammad Abdul Basit Malik, Matteo Leonetti, Gerard Canal, Martim Brandão
King's College London
London, UK
{zoe.a.evans, muhammad.a.malik, matteo.leonetti, gerard.canal, martim.brandao}@kcl.ac.uk

***Abstract*—Robot competitions such as RoboCup have long been a way promote and evaluate progress in robotics research. Since competitions have the potential to shape the direction of research, it is vital that they are inclusive of the people they will affect and critical of the practices and technologies they advocate for. This work aims to understand what gaps there are in robot competitions in regards to fostering inclusive development practices. In particular, we examine technical development reports from the 2024 Eindhoven RoboCup@Home league, and we find that many teams do not report on fairness or inclusivity practices, some subtask specifications are inherently problematic, and that relevant stakeholders are not involved in the design or evaluation of the competition. We offer recommendations to improve inclusivity in the RoboCup@Home league, which in turn could positively influence other areas of robotics development.**

***Index Terms*—inclusivity, human–robot interaction, competition, robotics, fairness**

I. Introduction

Robotics competitions contribute heavily to the field of human–robot–interaction, through establishing ways to test algorithms, set research directions and share advances in technology with the public. It has been argued that innovation competitions in general set the agenda in terms of what problems we consider worth solving, and this may have a subsequent impact on safety and inclusivity for affected groups who are not involved in this process [1].

This work focuses on RoboCup@Home, one of the leagues of the annual international RoboCup competition. This league has a focus on Human-Robot Interaction (HRI), with subgoals such as person detection, person following, interaction using speech, learning of faces, and learning of characteristics of a person.

In this paper we analyse inclusivity in the RoboCup@Home competition: including its development process (who participates in the choice of tasks), evaluation processes (who are the judges and what is the object of evaluation), problem framings (task definitions), team practices (algorithms, datasets and guardrails used), and team representation (participating group demographics and barriers towards participation). We conclude with recommendations towards more inclusive robot competitions.

II. Background and Related Work

Robot competitions have long been a way to practically realise some of the advancements in AI and robotics. Their focus on evaluating robots in the real world has allowed researchers to understand how this technology performs in brittler, more unpredictable situations [2]. Notable robot competitions include RoboCup [3], HuroCup [2] and the DARPA Robotics Challenge [4].

The focus in this paper is on RoboCup, a longstanding, global robotics competition launched in 1997. RoboCup is an attempt to promote AI and robotics research by providing a common task in which to evaluate theories, algorithms and methods - and is seen by some as the next best challenge for AI. In the year of its launch Kitano et al [5] proclaimed "because computer chess is about to complete its original goal, we need a new challenge, one that initiates a set of next-generation technologies. We believe that RoboCup fulfils such a demand". Originally launched as a robot–soccer competition, teams of competitors can now partake in leagues such as RoboCupSoccer, RoboCupRescue, RoboCup@Home, RoboCupIndustrial and RoboCupJunior [3]. Specifically, RoboCup@Home [6], with its strong emphasis on HRI in its tasks, will be the main focus of this paper.

Aside from being a useful benchmark, RoboCup has been regarded as an integral force in driving robotics research for the last 20 years. Tamburrini et al. discuss the framing of future goals and subgoals in RoboCup [7]. They argue that "The RoboCup long-term goal, it is surmised, enables one to shape a fruitful research agenda, insofar as it 'can easily create a series of well-directed subgoals' that are both feasible and technologically rewarding." Specifically for RoboCup@Home, they argue that the long term goal is to "develop autonomous robots sharing with humans the physical space of their homes and carrying out there a variety of useful domestic and assistive tasks."

However, other literature on innovation challenges for technological development casts doubt on this type of optimistic message. Sandvik, in Humanitarian Extractivism [1], argues that "innovation challenges and the solutions they offer can in themselves be a form of problem-framing." Competitions can shape the agenda of research. Sandvik argues that these types of competitions often exclude the groups they are attempting to aid; from problem design to the judging panel. This is in part because the barriers to entry in innovation competitions are high - and often the outcomes produced come from the reference point of the elite who are able to participate in them. A similar perspective could be taken with regards to robotics competitions, which is what this work aims to do.

This work was supported by the UKRI Centre for Doctoral Training in Safe and Trusted Artificial Intelligence [EP/S023356/1].

TABLE I
TEAM PRACTICES

Team ID	LLMs	CLIP	YOLO	Safety	Fairness	Transparency
1			✓	✓		
2	✓		✓			
3		✓		✓		
4			✓			
5			✓	✓		✓
6				✓		
7				✓		
8	✓		✓	✓		
9			✓	✓		
10	✓		✓	✓		
11	✓		✓			
12		✓				
13			✓			
14	✓		✓	✓		
15			✓	✓	✓	
16			✓	✓		
17	✓		✓	✓		
18	✓		✓			

III. TEAM PRACTICES

We start by investigating whether inclusivity is considered in teams' practices, in particular through the consideration of bias, and the development of guardrails against behaviour that can be considered unsafe or harmful to minority and socially-marginalized groups, such as the propagation of harmful stereotypes. Our aim is to investigate the degree to which inclusivity (in the sense of bias and safety around minority groups) is considered during the development or in the choice of out-of-the-box tools used by teams.

We collate the practices of the 18 qualified teams in the 2024 RoboCup@Home Open Platform league (OPL), as listed in Table I. We assign "IDs" to the teams instead of using their official team names, as our intent is not to judge or blame particular teams but to evaluate the community as a whole. The practices summarised in Table I are derived from the technical development reports (TDPs) submitted by qualified teams as part of the teams' application process for the competition. The table shows practices such as the use of specific algorithms known to lead to discriminatory and socially harmful behaviour, and whether team reports mention any ethical considerations or responsible-practices in their TDPs. We use principles such as safety, fairness, bias, inclusivity and explainability as categories of considerations, similar to other work on robot ethics [8], [9]. While it is possible that the teams did not report all design decisions or safeguarding processes, these reports are still a useful indicator of the degree of attention paid to aspects of bias and inclusivity, which can reinforce ideas of what are important problems.

One of the main computer vision algorithms used by teams is YOLO. YOLO [10] is a popular object detection and image segmentation model. 14 teams use YOLO for object recognition, a core part of many of the RoboCup@Home tasks, as well as for identifying people. *Only 1 out of 18 teams mentions fairness or bias considerations in their report.* However, YOLO is built on a variety of datasets known to be biased, such as COCO and ImageNet. For example, some versions of ImageNet contain offensive labels [11], and models trained on COCO have been shown to contain gender stereotypes, even when subsampled for balanced gender during training [12]. As shown by Wang et al [12], models trained with the COCO dataset have the potential to enforce gender stereotypes. For example, "cooking" was found to be more strongly associated with the label "woman" as opposed to "man". In a competition setting, therefore, when using YOLO or other COCO-based models, a robot may be more likely to retrieve a kitchen object from a woman than from man, due to gender being a factor that increases the likelihood of the cooking utensil label. Even though this type of biased behaviour may occur in the competition, team reports do not currently consider it.

Another vision-based algorithm mentioned in team TDPs is CLIP, which is used to infer the physical characteristics of volunteers. *None of the teams using CLIP mention a fairness or bias consideration in their report.* However, Hundt et al [13] has shown CLIP to produce harmful racist and sexist stereotypes about people when used in robotic applications. Often, in the competition, teams pass an image of a volunteer to CLIP, which returns a list of characteristics in text form. Therefore, if teams do not filter CLIP's output, and simply return the top characteristic provided, there is a risk that they inadvertently label (out-loud) a volunteer with a racist, sexist or stereotypical label based on their appearance. However, such risk is not acknowledged or considered.

Teams have also begun exploring using Large Language Models (LLMs) to plan robot actions during tasks. 6 teams in RoboCup@Home have now adopted the strategy of using LLMs to construct robot plans from verbal user input [14] [15] [16]. This technique is most often used for the General-Purpose-Service-Robot task (GPSR), where a robot must plan an unknown task from instructions given by a volunteer. *None of the teams using LLMs mention a fairness, inclusivity or bias consideration in their report.* However, investigations into LLMs for robot task planning by Azeem et al [17], found that LLMs are currently not robust enough to interact with a diverse range of people, often producing biased and unsafe outputs towards people with specific gender, nationality, ethnicity, age, and disability characteristics. For example, they found that LLMs are capable of producing plans for robots that involve removing a walking aid from a disabled person, and that always prefer to ask assistance to non-disabled users even if the type of disability is irrelevant to the task. In a competition setting, therefore, this could lead robots to ignore disabled users—a non-inclusive behaviour which is consistent with ableist microaggressions studied in the literature [17].

In general, when writing technical development reports, teams often did not discuss techniques around safeguarding, inclusion and ensuring unbiased outcomes from the datasets they use. Teams may not consider this part of their work as technically or scientifically interesting, or they may not have considered these issues. Teams have little incentive to focus on inclusion and bias when developing their robots, as the competition does not require this.

IV. Problem Framings

As expressed by Sandvik [1] in Humanitarian Extractivism, innovation competitions can be used to define the problems at hand that technology must solve (i.e. problem framing). Sandvik discusses “competitions for problem framing” in the context of humanitarian innovation contests. Innovation contests, removed from the people they aim to help, may inadvertently shape the agenda and idea of what problems should be solved. A similar argument can be applied to RoboCup; the choice of tasks and abilities required of the robot shape what is considered important research in the robotics and HRI community.

The RoboCup@Home rulebook for 2025 states that “the RoboCup@Home league aims to develop service and assistive robot technology with high relevance for future personal domestic applications” and that one of the core components of the competition is to produce “socially relevant results.” This mission statement enforces the vision that robots can and should be used in these ways in the future and that the resulting tasks and subtasks are desirable technologies. To that end, we will examine some of the tasks and subtasks set out by the rulebook for RoboCup@Home, and discuss how attempting to solve these tasks may result in harmful outcomes.

In the Receptionist Task, identifying three physical characteristics of a volunteer is a mandatory subtask [6]. The characteristics which should be identified are left to the team’s discretion, but gender and age are suitable features suggested by the rulebook. Some teams, for example, explicitly mention using computer vision to “estimate information such as age, gender, race, and emotions.” In fact, it has been suggested that future RoboCup@Home competitions should involve inferring “emotions and moods, activities, health and vital signs (inebriation, fatigue, sickness, sleep, etc.), skin and hair color, clothing names and styles, and identification by voice” [18].

Identifying physical characteristics is a clear subgoal for RoboCup and HRI, however, in recent years there has been opposition to this practice from other areas of the human-robot-interaction community [19]. It has been argued that using computer vision to detect sensitive characteristics such as race, gender, age, and culture, to name a few, has the potential to be harmful. For example, Williams et al. suggested that there are ontological, perceptual and deployment harms in using robots to detect race, gender and culture [19]. Misgendering (referring to someone with gendered terms that do not match their gender) is particularly harmful towards transgender individuals. Doing so within the competition can be harmful towards such individuals, as it “reinforces the idea that society doesn’t consider your gender ‘real”’ [20]. Additionally, even if all individuals participating in the competition with the robots identify as cis-gender, by encouraging gender classifications as part of the competition, we may inadvertadly be committing to a ’gender classification system’ that only realises a binary classification of gender [19]. This may have consequences to how participants, as future roboticists, develop robotics technology outside of the competitions.

TABLE II
RoboCup Tasks

Tasks	
Task	Focus
Help me carry	Person following, navigation in unmapped environments, social navigation.
General Purpose Service Robot	Task planning, object/people detection and recognition, object feature recognition, object manipulation
Receptionist	System Integration, Human-Robot Interaction, Person Detection, Person Recognition
Storing Groceries	Object detection and recognition, object feature recognition, object manipulation
Clean the Table	Object perception, manipulation in narrow spaces, and task planning.
Enhanced General Purpose Service Robot	Task planning, object/people detection and recognition, object feature recognition, object manipulation
Restuarant	Task planning, Online mapping, Navigation in unknown environments, Gesture detection, Verbal interaction and Object manipulation
Stickler for the Rules	Object perception, Human perception, Action recognition and Verbal interaction.

Inferring the emotions of volunteers from camera images also poses many ethical and moral questions. Emotion detection is not sensitive to cultural differences [21], and scientifically disputed [22] [23]. It is also sits at odds with human rights frameworks, with even the EU AI Act limiting the use of emotion recognition to only healthcare and safety contexts [21]. The use of emotion recognition is not directly encouraged by the RoboCup@Home rulebook, however neither is it discouraged. The harm here lies in the normalisation of these practices.

V. Development Processes

There are several committees in each RoboCup league that work towards the design, implementation and evaluation of the tasks in RoboCup. In RoboCup@Home, the executive committee are responsible for choosing the tasks in which teams can compete [6]. These tasks, and the main components they are aiming to evaluate, are highlighted in Table II.

The technical committee is responsible for implementing these tasks. They write the rulebook, and subsequently are responsible for acting as the referees during the competition. The technical committee is elected by the other teams participating in the competition and usually is composed of members of competing teams.

The technical committee is mostly composed of individuals with academic affiliations. For example, out of the 46 members of the technical committee of RoboCup@Home (from 2009 - 2024), according to personal webpages and LinkedIn profiles, 41 were academics or had affiliations associated with academia (universities, institutes, colleges).

No information is given in the RoboCup@Home documentation and rulebook on stakeholder input into the task choice or design. The rulebook states that one of the key criteria underpinning the RoboCup@Home competition is social relevance.

Examples of such socially relevant applications, according to the rulebook, are ‘a personal robot assistant, a guide robot for the blind, robot care for elderly people, and so forth.’ The creation of robots for disabled and older adult groups is a recurring argument for the relevance and importance of the RoboCup competition; however, there is little indication that members of these communities have been involved in the selection or design of the tasks. We believe this is a missed opportunity to truly design tasks that would fully encourage the development of technology that is useful for these communities.

VI. Evaluation Processes

The evaluation of teams’ performance is in the form of a series of trial runs at the task, where volunteers (volunteers from the local organisation) will interact with the robot. All tests are monitored by a referee, who is normally a member of the technical committee and will assign points. The robot’s performance is evaluated against the rulebook.

As in the development process, there is a lack of involvement from stakeholders, diverse affected communities and members of the public. It is hard to tell if a robot that may one day use these technologies for these communities will perform well, without these communities involved in the evaluation process.

VII. Team Representation

There is a substantial financial cost to attending RoboCup; the cost of research robots, travel, accommodation, registration, transportation of equipment, visas and more. This can be a significant barrier to entry for some prospective teams.

However, since debut of RoboCup@Home in 2006, 6 competitions have been held in Europe, 3 in the Americas (North and South), 5 in Asia, and 1 in Australia. In terms of the qualified teams (OPL), between 2007-2024 there have been *108* teams participating from Europe, *29* from North-America, *24* from South-America, *91* from Asia and *3* from Oceania(Australia) region. Europe had the highest number of teams in all editions except in those hosted in Thailand, Singapore and China (where Asian teams were the largest number). This shows the positive impact of current RoboCup practices in terms of fostering participation from diverse regions, towards fostering a globally inclusive community. We note, however, that hosting RoboCup in North/South-American regions did not lead to an increase in teams from those countries, which raises the question of other barriers to participation that are not just location (e.g. hardware and qualification barriers).

VIII. Discussion and Recommendations

In this paper we have argued that there are gaps in the RoboCup@Home competition regarding inclusivity and diversity. Namely, we showed that 1) teams often do not report any substantial action taken to mitigate bias and socially harmful outputs from the models they use, 2) the competition frames some of the tasks in problematic (uninclusive) ways, 3) the development and evaluation processes do not include stakeholders to guarantee the relevance of and feedback on the tasks.

RoboCup@Home has achieved much in international cooperation; hosting the competition in multiple parts of the world since its conception. To further these achievements, we offer realisable recommendations to improve inclusivity, and safety for diverse groups, in the RoboCup@Home competition.

Our recommendations are as follows.

1) Social and psychological safety checks should be added to preconditions or scoring methods of competitions. Since passing physical safety checks is often a non-negotiable condition of participating in the rest of the competition, similar guarantees for the social and psychological safety of volunteers and participants should also be a requirement. These could include showing (via report or demonstration) that harmful statements (e.g. misgendering, stereotyping, or stating personal features in a harmful way) and the underperformance or ignoring for specific groups, cannot occur during the robots execution of the task. We should aim to use datasets that detail their design choices, requirements, and evaluation methods [24].
2) Further explicit guidance should be written into the rulebook over what kind of labels, features and characteristics teams should aim for their robot to detect. A wider (and global) discussion should be had about what norms and values we propagate with the labels we proscribe to volunteers during tasks, and be explicit about what we would like to encourage and avoid.
3) Tasks should include subtasks that focus on ethical decision-making. For example, in the General-Purpose-Service-Robot task, bonus points could be allocated for guard-railing against unethical or unsafe commands. Robot competitions that include ethical-decision making have already been trialled [25]. This could be used to emphasize robot ethics as an important area of research and as an interesting technical challenge.
4) Members of diverse affected groups—such as representatives of elder care homes, older adult, disability and anti-racism NGOs—should be more present in the design and judgement of competitions that affect them, as has been shown from other competition literature [1].

IX. Conclusion

Robot competitions have the potential to shape the research directions in the field of robotics. For this reason, its important that competitions are inclusive to all types of people they will potentially impact. Here, we analysed inclusivity practices in a major robot competition league, RoboCup@Home. We argued that, from the Rulebook and teams technical development reports, there are currently gaps in inclusivity-related practices—and we suggested recommendations to remedy these gaps.

References

[1] Kristin Bergtora Sandvik. *Humanitarian extractivism The digital transformation of aid.* Manchester University Press, October 2023.

[2] John Anderson, Jacky Baltes, and Chi Tai Cheng. Robotics competitions as benchmarks for AI research. *The Knowledge Engineering Review*, 26(1):11–17, February 2011.

[3] RoboCup Federation, Jul 2023.

[4] Eric Krotkov, Douglas Hackett, Larry Jackel, Michael Perschbacher, James Pippine, Jesse Strauss, Gill Pratt, and Christopher Orlowski. The darpa robotics challenge finals: Results and perspectives. *The DARPA robotics challenge finals: Humanoid robots to the rescue*, pages 1–26, 2018.

[5] Hiroaki Kitano, Minoru Asada, Yasuo Kuniyoshi, Itsuki Noda, Eiichi Osawa, and Hitoshi Matsubara. RoboCup: A Challenge Problem for AI. 18(1):73–73.

[6] Justin Hart, Alexander Moriarty, Katarzyna Pasternak, Johannes Kummert, Alina Hawkin, Vanessa Hassouna, Juan Diego Pena Narvaez, Leroy Ruegemer, Leander von Seelstrang, Peter Van Dooren, Juan Jose Garcia, Akinobu Mitzutani, Yuqian Jiang, Tatsuya Matsushima, and Riccardo Polvara. Robocup@home 2024: Rules and regulations. https://github.com/RoboCupAtHome/RuleBook/releases/tag/2024.1, 2024.

[7] Guglielmo Tamburrini. On the ethical framing of research programs in robotics. 31(4):463–471.

[8] Sofya Langman, Nicole Capicotto, Yaser Maddahi, and Kourosh Zareinia. Roboethics principles and policies in europe and north america. *SN Applied Sciences*, 3(12):857, 2021.

[9] High-Level Expert Group on Artificial Intelligence. Ethics guidelines for trustworthy ai, April 2019. Accessed on January 24, 2025.

[10] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 779–788, 2016.

[11] Kaiyu Yang, Klint Qinami, Li Fei-Fei, Jia Deng, and Olga Russakovsky. Towards Fairer Datasets: Filtering and Balancing the Distribution of the People Subtree in the ImageNet Hierarchy. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 547–558.

[12] Tianlu Wang, Jieyu Zhao, Mark Yatskar, Kai-Wei Chang, and Vicente Ordonez. Balanced Datasets Are Not Enough: Estimating and Mitigating Gender Bias in Deep Image Representations.

[13] Andrew Hundt, William Agnew, Vicky Zeng, Severin Kacianka, and Matthew Gombolay. Robots Enact Malignant Stereotypes. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, pages 743–756. Association for Computing Machinery.

[14] Ruoyu Wang, Zhipeng Yang, Zinan Zhao, Xinyan Tong, Zhi Hong, and Kun Qian. LLM-based Robot Task Planning with Exceptional Handling for General Purpose Service Robots, May 2024.

[15] Taewoong Kang, Joonyoung Kim, Jaebong Yi, Shady Nasrat, Min-Seong Jo, Min-Seong Jo, Jae-Moon Park, Tae-Young Kim, Byeong-Ju Kim, Seung-Joon Yi, and Byoung-Tak Zhang. Tidyboy-OPL: RoboCup@Home Open Platform League Team Description Paper.

[16] Liman Wang and Hanyang Zhong. LLM-SAP: Large Language Models Situational Awareness Based Planning.

[17] Rumaisa Azeem, Andrew Hundt, Masoumeh Mansouri, and Martim Brandão. LLM-Driven Robots Risk Enacting Discrimination, Violence, and Unlawful Actions.

[18] Mauricio MATAMOROS, Viktor SEIB, and Dietrich PAULUS. Trends, Challenges and Adopted Strategies in RoboCup@Home. In *2019 IEEE International Conference on Autonomous Robot Systems and Competitions (ICARSC)*, pages 1–6.

[19] Tom Williams. The Eye of the Robot Beholder: Ethical Risks of Representation, Recognition, and Reasoning over Identity Characteristics in Human-Robot Interaction. In *Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '23, pages 1–10. Association for Computing Machinery.

[20] Os Keyes. The Misgendering Machines: Trans/HCI Implications of Automatic Gender Recognition. 2:88:1–88:22.

[21] Martina Mattioli and Federico Cabitza. Not in My Face: Challenges and Ethical Considerations in Automatic Face Emotion Recognition Technology. 6(4):2201–2231.

[22] Nicola Binetti, Nadejda Roubtsova, Christina Carlisi, Darren Cosker, Essi Viding, and Isabelle Mareschal. Genetic algorithms reveal profound individual differences in emotion recognition. 119(45):e2201380119.

[23] Juan I. Durán and José-Miguel Fernández-Dols. Do emotions result in their predicted facial expressions? A meta-analysis of studies on the co-occurrence of expression and emotion. 21(7):1550–1569.

[24] Romain Egele, Julio Junior, CS Jacques, Jan N van Rijn, Isabelle Guyon, Xavier Baró, Albert Clapés, Prasanna Balaprakash, Sergio Escalera, Thomas Moeslund, et al. Ai competitions and benchmarks: Dataset development. *arXiv preprint arXiv:2404.09703*, 2024.

[25] Jimin Rhim, Cheng Lin, Alexander Werner, Brandon DeHart, Vivian Qiang, Shalaleh Rismani, and given-i=Aj family=Moon, given=Ajung. Roboethics as a Design Challenge: Lessons Learned from the Roboethics to Design and Development Competition. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 11244–11250.