# ARTIFICIAL INTELLIGENCE APPLICATIONS IN ROBOTICS: FROM VISION TO AUTONOMY

**Anup Sanjay Jadhao**
*Research Scholar, Shri Shivaji college of arts,Commerce and science, Akola*
*jadhao.anups@gmail.com*

**Dr. S. M. Chavan**
*Assistant Professor, Shri Shivaji college of arts, Commerce and science, Akola*
*santoshchavan9881@gmail.com*

**Abstract**
*Artificial Intelligence (AI) is reshaping robotics by moving beyond rule-based programming toward adaptive and embodied intelligence. Foundation models (2020–2025) play a central role, enabling robots to integrate perception, reasoning, and action across varied tasks and environments. Studies emphasize their strengths in zero-shot generalization, multimodal learning, and human–robot interaction, alongside challenges in data efficiency, robustness, safety, and ethics. Practical advances, including RT-1 for large-scale control, vision–language models for imitation learning, RoboCat for continual self-improvement, and RobotxR1 for reinforcement-driven reasoning, illustrate their potential to enhance autonomy and adaptability. Complementary work in deep reinforcement learning further supports real-time navigation and control. Overall, foundation models provide a transformative pathway toward scalable, flexible, and intelligent robotic systems, though significant technical and ethical barriers remain. In this paper, we focused on the role of foundation models in transforming robotics.*

***Keywords:*** *Artificial Intelligence (AI), Robotics, Models, Deep Learning, Autonomy, Generalization, Adaptability.*

## 1. Introduction:

Artificial Intelligence (AI) is rapidly transforming robotics, moving from rule-based machines to adaptive, general-purpose, and autonomous agents. While traditional robotics was limited by rigid programming, the rise of foundation models (2020–2025) has enabled robots to perceive, reason, and act with far greater flexibility. These advances bridge the gap between computer vision–based perception and autonomous decision-making, allowing robots to adapt in dynamic, unstructured environments. Surveys highlight the central role of foundation models in robotic intelligence. For example, some studies emphasize that pre-trained large-scale models can act as general backbones for perception and control, reducing dependence on narrow, task-specific methods [1]. Others review their applications in manipulation, navigation, and human-robot interaction, while also noting challenges such as data efficiency, robustness, and safety [2] [3].

Practical breakthroughs further show their potential. RT-1 introduced large-scale control using transformer architectures [4], while vision-language foundation models proved effective for imitation learning from demonstrations [5]. Similarly, vision-language-action flow models enable generalization across tasks by integrating perception, language, and motor control [6]. Self-improving and reinforcement-driven approaches extend autonomy even further. RoboCat verified continual self-improvement through reskilling on new data [7], while RobotxR1 combined large language models with strengthening learning for closed-loop reasoning [8].

At the task level, deep RL for vision-based obstacle avoidance shows how robots can achieve real-time adaptive navigation [9]. Together, these works suggest that integrating vision, language, and action through foundation models is accelerating the move toward embodied AI robots that learn, perceive, and act in ways similar to human cognition [10].

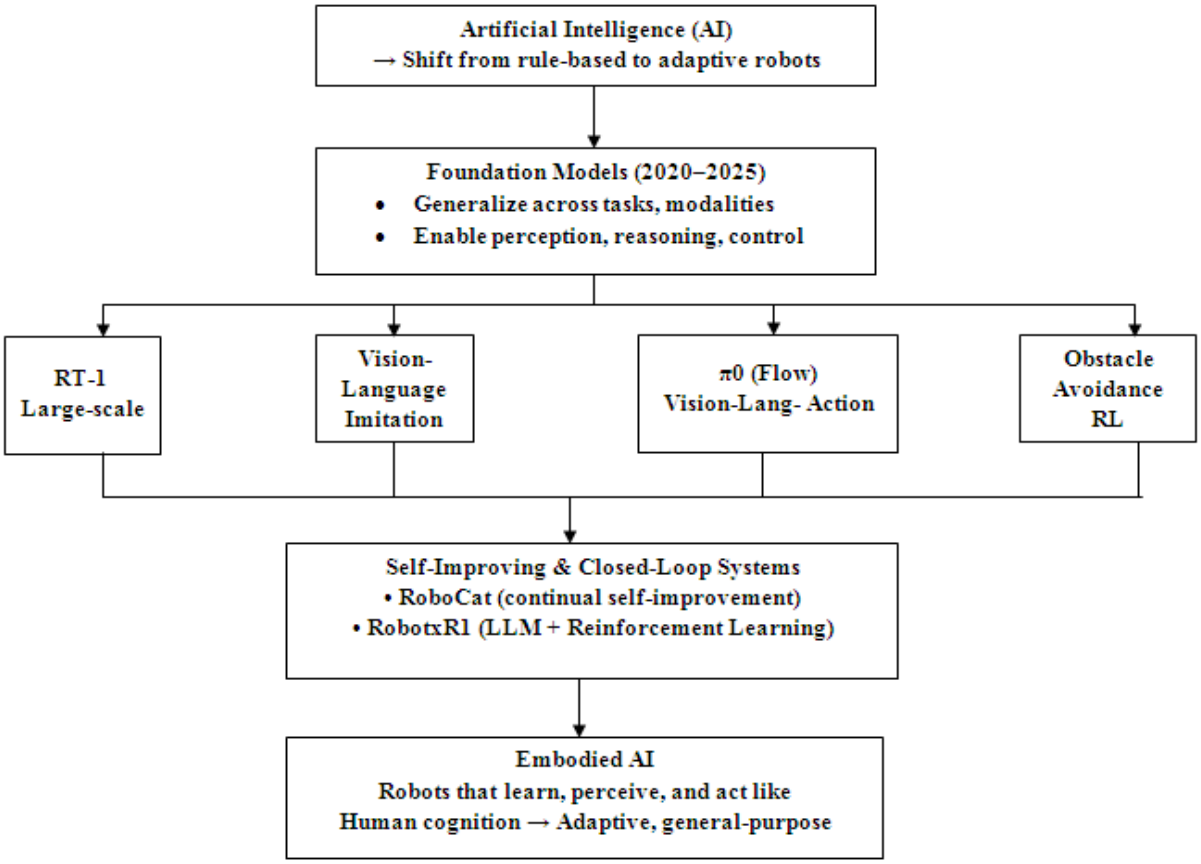


Fig. 1 from AI to Embodied AI

## 2. Literature review:

Hu et al. (2023) provided one of the most comprehensive examinations of the role of foundation models in advancing general-purpose robotics, situating robotics at the intersection of machine learning, computer vision, natural language processing, and control. The authors conduct both a survey and a meta-analysis, offering an in-depth review of how large-scale pre trained models, particularly vision-language and multimodal transformers, are increasingly leveraged to enhance perception, planning, and action in robotic systems. By synthesizing existing research, they identify the strengths of foundation models in enabling zero-shot generalization, multimodal reasoning, and embodied intelligence, while also pointing to persistent challenges such as data efficiency, robustness in dynamic environments, and alignment between learned representations and real-world tasks. Their meta-analysis underscores the transformative potential of foundation models for creating scalable robotic systems capable of autonomous and adaptive behaviour across varied domains, thereby positioning such models as a critical step toward embodied artificial intelligence. This work is widely regarded as a landmark contribution in establishing the conceptual and methodological foundations for general-purpose robotics research. [1]

Firoozi et al. (2023) present a systematic exploration of the applications, challenges, and future directions of foundation models in robotics, positioning them as a transformative paradigm for advancing robot autonomy and adaptability. The paper emphasizes how large-scale pre trained models, particularly those integrating vision, language, and action modalities, can significantly enhance robotic perception, planning, manipulation, and human-robot interaction. Unlike traditional task-specific methods, foundation models enable robots to generalize across tasks, environments, and sensory inputs, making them more scalable and versatile. However, the authors also highlight critical limitations, including the high computational and data requirements, safety and reliability concerns in real-world deployments, and the pressing need for ethical frameworks to govern their use. Their forward-looking perspective underscores the importance of interdisciplinary research to overcome these barriers, while pointing to opportunities in leveraging foundation models for embodied intelligence and human centred robotics. This work thus serves as both a roadmap

and a cautionary note for integrating foundation models into the next generation of robotic systems. [2]

Xu and Zhao (2023) provide a focused examination of best practices in applying foundation models to robotics, emphasizing methodological rigor and practical considerations for effective deployment. Their work highlights how the integration of large-scale pre trained models into robotic systems requires careful handling of data collection, model fine-tuning, and domain adaptation to ensure robust generalization in real-world scenarios. They stress the importance of leveraging multimodal inputs—such as vision, language, and sixth sense while maintaining efficiency through techniques like parameter-efficient training and modular architectures. Unlike broader surveys, this study adopts a prescriptive stance by outlining guidelines for aligning foundation model capabilities with robotics-specific tasks, thereby addressing challenges such as safety, interpretability, and hardware limitations. By systematizing lessons learned from recent advances, Xu and Zhao contribute a practice-oriented perspective that complements more theoretical or application-driven studies, offering researchers and practitioners a framework to bridge the gap between foundation model theory and embodied robotic intelligence. [3]

Brohan et al. (2022) present RT-1, a robotics transformer model designed to unify robotic control through large-scale learning. The model leverages a transformer-based architecture trained on extensive real-world demonstrations, enabling strong generalization across tasks and instructions. The study highlights how scaling data and model capacity enhances performance in long-horizon and compositional tasks, emphasizing the role of natural language conditioning and multimodal integration. RT-1 demonstrates robustness and adaptability, marking a significant step toward generalist robotic agents and establishing a foundation for future advancements in real-world embodied intelligence. [4]

Li et al. (2023) investigate the use of vision-language foundation models (VLFMs) as robot imitators, focusing on their ability to generalize across tasks through multimodal learning. The study emphasizes how VLFMs leverage large-scale pre training to map visual and textual inputs into meaningful action representations, enhancing imitation learning efficiency. By aligning perception and instruction understanding, these models show promise in bridging the gap between human demonstrations and robotic execution. The work highlights their potential in enabling scalable, flexible, and adaptable robot learning while also addressing challenges such as grounding, robustness, and deployment in real-world settings. [5]

Black et al. (2023) introduce π0: a vision-language-action (VLA) flow model designed for general robot control, positioning it as a step toward unified embodied intelligence. The study emphasizes how π0 integrates multimodal data visual perception, natural language, and action flow to enable robots to interpret instructions and perform tasks in a more generalized manner. By aligning language with perception and control, the model aims to overcome limitations of task-specific approaches, offering a scalable pathway for robust real-world deployment. The work highlights its significance in advancing foundation models for robotics while identifying ongoing challenges in grounding, adaptability, and safe deployment. [6]

Bousmalis et al. (2023) presented RoboCat, a self-improving generalist agent for robotic management that forces the scaling properties of foundation models. The model is trained on diverse multimodal datasets and designed to adapt to new tasks with minimal additional data, thereby demonstrating strong generalization capabilities. A key contribution is its ability to self-improve through iterative data collection and fine-tuning, allowing the system to continually expand its skill set across a wide range of robotic manipulation tasks. This approach highlights the potential of combining large-scale pre training with autonomous adaptation, setting a precedent for more versatile and scalable robotic learning systems. [7]

Boyle et al. (2023) introduce RobotxR1, a framework that integrates large language models (LLMs) with robotics through closed-loop reinforcement learning to enable embodied intelligence. Unlike traditional static prompt-based approaches, RobotxR1 emphasizes continuous feedback and iterative learning, allowing robots to refine decision-making and improve task performance dynamically. The study highlights the role of reinforcement signals in bridging high-level natural language understanding with low-level robotic control, thereby enhancing adaptability in real-world environments. This work underscores the growing importance of combining LLMs with reinforcement learning to achieve more autonomous, general-purpose robotic agents. [8]

Wenzel et al. (2020) present a study on vision-based obstacle avoidance in mobile robotics using deep reinforcement learning (DRL). Their work departs from reliance on handcrafted features or traditional path-planning methods by leveraging

raw visual input to train agents for navigation in complex environments. The research demonstrates how DRL enables robots to learn robust obstacle-avoidance policies that generalize across dynamic and cluttered scenarios. By integrating deep learning with reinforcement-based decision-making, the study highlights the potential of DRL to improve autonomy in mobile robots, particularly in environments where explicit modelling of obstacles is difficult or infeasible. [9]

Xu et al. (2023) provide a comprehensive survey on the integration of foundation models into robotics, emphasizing their role in advancing embodied AI. The paper categorizes existing approaches based on perception, decision-making, and control, showing how foundation models can unify these traditionally separate components. The authors highlight the scalability and adaptability of such models, particularly in enabling robots to handle diverse tasks without extensive task-specific retraining. They also discuss key challenges, including high computational demands, the need for domain adaptation to physical environments, and safety concerns in deployment. Importantly, the study underscores the promise of foundation models in bridging the gap between simulation and real-world robotics, setting a research agenda for more general-purpose, intelligent, and embodied robotic systems. [10]

## 3. Research Work:

Foundation models are changing robotics by helping robots learn and adapt too many different tasks, inputs, and environments. Surveys [1][2][3][10] show that these models are strong in seeing, planning, and controlling actions, but also face challenges like needing lots of data, handling changes in the real world, staying safe, and following ethics. Key progress includes RT-1 for large-scale robot control [4], VLFMs for learning from human examples [5], π0: for linking vision, language, and actions [6], RoboCat as a self-learning robot agent [7], and RobotxR1 that connects language models with reinforcement learning [8]. Research on deep reinforcement learning for avoiding obstacles [9] also adds support. Overall, these works show that foundation models can lead to smarter, more flexible, and more independent robots.

## 4. Conclusion:

Foundation models are transforming robotics by enabling robots to learn, adapt, and perform a wide range of tasks across different environments. Unlike traditional rule-based approaches, these models allow robots to combine perception, planning, and action in a more flexible and intelligent way. Recent developments such as large-scale control systems, vision-language models, self-improving agents, and reinforcement learning techniques have shown how robots can generalize skills, learn from demonstrations, and continuously improve their abilities. While challenges like data needs, robustness, safety, and ethical concerns remain, foundation models offer a promising pathway toward creating truly intelligent, autonomous, and adaptive robotic systems that can function effectively in real-world, dynamic situations.

## 5. References:

1. Hu, Y., Xie, Q., Jain, V., Francis, J., Patrikar, J., Keetha, N., Kim, S., Xie, Y., Zhang, T., Fang, H.-S., Zhao, S., Omidshafiei, S., Kim, D.-K., Agha-mohammadi, A. Sycara, K., Johnson-Roberson, M., Batra, D., Wang, X., Scherer, S., Wang, C., Kira, Z., Xia, F., & Bisk, Y. (2023). Toward general-purpose robots via foundation models: A survey and meta-analysis.
2. Firoozi, R., Tucker, J., Tian, S., Majumdar, A., Sun, J., Liu, W., Zhu, Y., Song, S., Kapoor, A., Hausman, K., Ichter, B., Driess, D., Wu, J., Lu, C., & Schwager, M. (2023). Foundation models in robotics: Applications, challenges, and the future.
3. Xu, S., & Zhao, H. (2023). Foundation models for robotics: Best known practices.
4. Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Jasmine Hsu, Julian Ibarz, Brian Ichter, Alex Irpan, Tomas Jackson, Sally Jesmonth, Nikhil J Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Isabel Leal, Kuang-Huei Lee, Sergey Levine, Yao Lu, Utsav Malla, Deeksha Manjunath, Igor Mordatch, Ofir Nachum, Carolina Parada, Jodilyn Peralta, Emily Perez, Karl Pertsch, Jornell Quiambao, Kanishka Rao, Michael Ryoo, Grecia Salazar, Pannag Sanketi, Kevin Sayed, Jaspiar Singh, Sumedh Sontakke, Austin Stone, Clayton Tan, Huong Tran, Vincent Vanhoucke, Steve Vega, Quan Vuong, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Tianhe Yu, Brianna Zitkovich (2022). RT-1: Robotics transformer for real-world control at scale.
5. Li, X., Liu, M., Zhang, H., Yu, C., Xu, J., Wu, H., Cheang, C., Jing, Y., Zhang, W., Liu, H., Li, H., & Kong, T. (2023). Vision-language foundation models as effective robot imitators.

6. Black, K., Brown, N., Driess, D., Esmail, A., Equi, M., Finn, C., Fusai, N., Groom, L., Hausman, K., Ichter, B., Jakubczak, S., Jones, T., Ke, L., Levine, S., Li-Bell, A., Mothukuri, M., Nair, S., Pertsch, K., Shi, L. X., Tanner, J., Vuong, Q., Walling, A., Wang, H., & Zhilinsky, U. (2023). π0: A vision-language-action flow model for general robot control.
7. Konstantinos Bousmalis, Giulia Vezzani, Dushyant Rao, Coline Devin, Alex X. Lee, Maria Bauza, Todor Davchev, Yuxiang Zhou, Agrim Gupta, Akhil Raju , Antoine Laurens, Claudio Fantacci, Valentin Dalibard, Martina Zambelli, Murilo F. Martins, Rugile Pevceviciute, Michiel Blokzijl, Misha Denil, Nathan Batchelor, Thomas Lampe, Emilio Parisotto, Konrad Żołna, Scott Reed, Sergio Gómez Colmenarejo, Jon Scholz, Abbas Abdolmaleki, Oliver Groth, Jean-Baptiste Regli, Oleg Sushkov, Tom Rothörl, José Enrique Chen, Yusuf Aytar, Dave Barker, Joy Ortiz, Martin Riedmiller, Jost Tobias Springenberg, Raia Hadsell, Francesco Nori, and Nicolas Heess. (2023). RoboCat: A self-improving generalist agent for robotic manipulation.
8. Boyle, L., Baumann, N., Sivasothilingam, P., Magno, M., & Benini, L. (2023). RobotxR1: Enabling embodied robotic intelligence on large language models through closed-loop reinforcement learning.
9. Wenzel, P., Schön, T., Leal-Taixé, L., & Cremers, D. (2020). Vision-based mobile robotics obstacle avoidance with deep reinforcement learning.
10. Xu, Z., Wu, K., Wen, J., Li, J., Liu, N., Che, Z., & Tang, J. (2023). A survey on robotics with foundation models: Toward embodied AI.