**Chapter 13\_ Human-in-the-Loop**

Chapter 13: Human-in-the-Loop

The Human-in-the-Loop (HITL) pattern represents a pivotal strategy in the development and deployment of Agents. It deliberately interweaves the unique strengths of human cognition—such as judgment, creativity, and nuanced understanding—with the computational power and efficiency of AI. This strategic integration is not merely an option but often a necessity, especially as AI systems become increasingly embedded in critical decision-making processes.

The core principle of HITL is to ensure that AI operates within ethical boundaries, adheres to safety protocols, and achieves its objectives with optimal effectiveness. These concerns are particularly acute in domains characterized by complexity, ambiguity, or significant risk, where the implications of AI errors or misinterpretations can be substantial. In such scenarios, full autonomy—where AI systems function independently without any human intervention—may prove to be imprudent. HITL acknowledges this reality and emphasizes that even with rapidly advancing AI technologies, human oversight, strategic input, and collaborative interactions remain indispensable.

The HITL approach fundamentally revolves around the idea of synergy between artificial and human intelligence. Rather than viewing AI as a replacement for human workers, HITL positions AI as a tool that augments and enhances human capabilities. This augmentation can take various forms, from automating routine tasks to providing data-driven insights that inform human decisions. The end goal is to create a collaborative ecosystem where both humans and AI Agents can leverage their distinct strengths to achieve outcomes that neither could accomplish alone.

In practice, HITL can be implemented in diverse ways. One common approach involves humans acting as validators or reviewers, examining AI outputs to ensure accuracy and identify potential errors. Another implementation involves humans actively guiding AI behavior, providing feedback or making corrections in real-time. In more complex setups, humans may collaborate with AI as partners, jointly solving problems or making decisions through interactive dialog or shared interfaces. Regardless of the specific implementation, the HITL pattern underscores the importance of maintaining human control and oversight, ensuring that AI systems remain aligned with human ethics, values, goals, and societal expectations.

**Human-in-the-Loop Pattern Overview**

The Human-in-the-Loop (HITL) pattern integrates artificial intelligence with human input to enhance Agent capabilities. This approach acknowledges that optimal AI performance frequently requires a combination of automated processing and human insight, especially in scenarios with high complexity or ethical considerations. Rather than replacing human input, HITL aims to augment human abilities by ensuring that critical judgments and decisions are informed by human understanding.

HITL encompasses several key aspects: Human Oversight, which involves monitoring AI agent performance and output (e.g., via log reviews or real-time dashboards) to ensure adherence to guidelines and prevent undesirable outcomes. Intervention and Correction occurs when an AI agent encounters errors or ambiguous scenarios and may request human intervention; human operators can rectify errors, supply missing data, or guide the agent, which also informs future agent improvements. Human Feedback for Learning is collected and used to refine AI models, prominently in methodologies like reinforcement learning with human feedback, where human preferences directly influence the agent's learning trajectory. Decision Augmentation is where an AI agent provides analyses and recommendations to a human, who then makes the final decision, enhancing human decision-making through AI-generated insights rather than full autonomy. Human-Agent Collaboration is a cooperative interaction where humans and AI agents contribute their respective strengths; routine data processing may be handled by the agent, while creative problem-solving or complex negotiations are managed by the human. Finally, Escalation Policies are established protocols that dictate when and how an agent should escalate tasks to human operators, preventing errors in situations beyond the agent's capability.

Implementing HITL patterns enables the use of Agents in sensitive sectors where full autonomy is not feasible or permitted. It also provides a mechanism for ongoing improvement through feedback loops. For example, in finance, the final approval of a large corporate loan requires a human loan officer to assess qualitative factors like leadership character. Similarly, in the legal field, core principles of justice and accountability demand that a human judge retain final authority over critical decisions like sentencing, which involve complex moral reasoning.

**Caveats: Despite its benefits, the HITL pattern has significant caveats, chief among them being a lack of scalability. While human oversight provides high accuracy, operators cannot manage millions of tasks, creating a fundamental trade-off that often requires a hybrid approach combining automation for scale and HITL for accuracy. Furthermore, the effectiveness of this pattern is heavily dependent on the expertise of the human operators; for example, while an AI can generate software code, only a skilled developer can accurately identify subtle errors and provide the correct guidance to fix them. This need for expertise also applies when using HITL to generate training data, as human annotators may require special training to learn how to correct an AI in a way that produces high-quality data. Lastly, implementing HITL raises significant privacy concerns, as sensitive information must often be rigorously anonymized before it can be exposed to a human operator, adding another layer of process complexity.**

**Practical Applications & Use Cases**

The Human-in-the-Loop pattern is vital across a wide range of industries and applications, particularly where accuracy, safety, ethics, or nuanced understanding are paramount.

* **Content Moderation:** AI agents can rapidly filter vast amounts of online content for violations (e.g., hate speech, spam). However, ambiguous cases or borderline content are escalated to human moderators for review and final decision, ensuring nuanced judgment and adherence to complex policies.
* **Autonomous Driving:** While self-driving cars handle most driving tasks autonomously, they are designed to hand over control to a human driver in complex, unpredictable, or dangerous situations that the AI cannot confidently navigate (e.g., extreme weather, unusual road conditions).
* **Financial Fraud Detection:** AI systems can flag suspicious transactions based on patterns. However, high-risk or ambiguous alerts are often sent to human analysts who investigate further, contact customers, and make the final determination on whether a transaction is fraudulent.
* **Legal Document Review:** AI can quickly scan and categorize thousands of legal documents to identify relevant clauses or evidence. Human legal professionals then review the AI's findings for accuracy, context, and legal implications, especially for critical cases.
* **Customer Support (Complex Queries):** A chatbot might handle routine customer inquiries. If the user's problem is too complex, emotionally charged, or requires empathy that the AI cannot provide, the conversation is seamlessly handed over to a human support agent.
* **Data Labeling and Annotation:** AI models often require large datasets of labeled data for training. Humans are put in the loop to accurately label images, text, or audio, providing the ground truth that the AI learns from. This is a continuous process as models evolve.
* **Generative AI Refinement:** When an LLM generates creative content (e.g., marketing copy, design ideas), human editors or designers review and refine the output, ensuring it meets brand guidelines, resonates with the target audience, and maintains quality.
* **Autonomous Networks:** AI systems are capable of analyzing alerts and forecasting network issues and traffic anomalies by leveraging key performance indicators (KPIs) and identified patterns. Nevertheless, crucial decisions—such as addressing high-risk alerts—are frequently escalated to human analysts. These analysts conduct further investigation and make the ultimate determination regarding the approval of network changes.

This pattern exemplifies a practical method for AI implementation. It harnesses AI for enhanced scalability and efficiency, while maintaining human oversight to ensure quality, safety, and ethical compliance.

"Human-on-the-loop" is a variation of this pattern where human experts define the overarching policy, and the AI then handles immediate actions to ensure compliance. Let's consider two examples:

* **Automated financial trading system**: In this scenario, a human financial expert sets the overarching investment strategy and rules. For instance, the human might define the policy as: "Maintain a portfolio of 70% tech stocks and 30% bonds, do not invest more than 5% in any single company, and automatically sell any stock that falls 10% below its purchase price." The AI then monitors the stock market in real-time, executing trades instantly when these predefined conditions are met. The AI is handling the immediate, high-speed actions based on the slower, more strategic policy set by the human operator.
* **Modern call center**: In this setup, a human manager establishes high-level policies for customer interactions. For instance, the manager might set rules such as "any call mentioning 'service outage' should be immediately routed to a technical support specialist," or "if a customer's tone of voice indicates high frustration, the system should offer to connect them directly to a human agent." The AI system then handles the initial customer interactions, listening to and interpreting their needs in real-time. It autonomously executes the manager's policies by instantly routing the calls or offering escalations without needing human intervention for each individual case. This allows the AI to manage the high volume of immediate actions according to the slower, strategic guidance provided by the human operator.

**Hands-On Code Example**

To demonstrate the Human-in-the-Loop pattern, an ADK agent can identify scenarios requiring human review and initiate an escalation process . This allows for human intervention in situations where the agent's autonomous decision-making capabilities are limited or when complex judgments are required. This is not an isolated feature; other popular frameworks have adopted similar capabilities. LangChain, for instance, also provides tools to implement these types of interactions.

|  |
| --- |
| from google.adk.agents import Agent  from google.adk.tools.tool\_context import ToolContext  from google.adk.callbacks import CallbackContext  from google.adk.models.llm import LlmRequest  from google.genai import types  from typing import Optional  # Placeholder for tools (replace with actual implementations if needed)  def troubleshoot\_issue(issue: str) -> dict:  return {"status": "success", "report": f"Troubleshooting steps for {issue}."}  def create\_ticket(issue\_type: str, details: str) -> dict:  return {"status": "success", "ticket\_id": "TICKET123"}  def escalate\_to\_human(issue\_type: str) -> dict:  # This would typically transfer to a human queue in a real system  return {"status": "success", "message": f"Escalated {issue\_type} to a human specialist."}  technical\_support\_agent = Agent(  name="technical\_support\_specialist",  model="gemini-2.0-flash-exp",  instruction="""  You are a technical support specialist for our electronics company.  FIRST, check if the user has a support history in state["customer\_info"]["support\_history"]. If they do, reference this history in your responses.  For technical issues:  1. Use the troubleshoot\_issue tool to analyze the problem.  2. Guide the user through basic troubleshooting steps.  3. If the issue persists, use create\_ticket to log the issue.  For complex issues beyond basic troubleshooting:  1. Use escalate\_to\_human to transfer to a human specialist.  Maintain a professional but empathetic tone. Acknowledge the frustration technical issues can cause, while providing clear steps toward resolution.  """,  tools=[troubleshoot\_issue, create\_ticket, escalate\_to\_human]  )  def personalization\_callback(  callback\_context: CallbackContext, llm\_request: LlmRequest  ) -> Optional[LlmRequest]:  """Adds personalization information to the LLM request."""  # Get customer info from state  customer\_info = callback\_context.state.get("customer\_info")  if customer\_info:  customer\_name = customer\_info.get("name", "valued customer")  customer\_tier = customer\_info.get("tier", "standard")  recent\_purchases = customer\_info.get("recent\_purchases", [])  personalization\_note = (  f"\nIMPORTANT PERSONALIZATION:\n"  f"Customer Name: {customer\_name}\n"  f"Customer Tier: {customer\_tier}\n"  )  if recent\_purchases:  personalization\_note += f"Recent Purchases: {', '.join(recent\_purchases)}\n"  if llm\_request.contents:  # Add as a system message before the first content  system\_content = types.Content(  role="system", parts=[types.Part(text=personalization\_note)]  )  llm\_request.contents.insert(0, system\_content)  return None # Return None to continue with the modified request |

This code offers a blueprint for creating a technical support agent using Google's ADK, designed around a HITL framework. The agent acts as an intelligent first line of support, configured with specific instructions and equipped with tools like troubleshoot\_issue, create\_ticket, and escalate\_to\_human to manage a complete support workflow. The escalation tool is a core part of the HITL design, ensuring complex or sensitive cases are passed to human specialists.

A key feature of this architecture is its capacity for deep personalization, achieved through a dedicated callback function. Before contacting the LLM, this function dynamically retrieves customer-specific data—such as their name, tier, and purchase history—from the agent's state. This context is then injected into the prompt as a system message, enabling the agent to provide highly tailored and informed responses that reference the user's history. By combining a structured workflow with essential human oversight and dynamic personalization, this code serves as a practical example of how the ADK facilitates the development of sophisticated and robust AI support solutions.

**At Glance**

**What:** AI systems, including advanced LLMs, often struggle with tasks that require nuanced judgment, ethical reasoning, or a deep understanding of complex, ambiguous contexts. Deploying fully autonomous AI in high-stakes environments carries significant risks, as errors can lead to severe safety, financial, or ethical consequences. These systems lack the inherent creativity and common-sense reasoning that humans possess. Consequently, relying solely on automation in critical decision-making processes is often imprudent and can undermine the system's overall effectiveness and trustworthiness.

**Why:** The Human-in-the-Loop (HITL) pattern provides a standardized solution by strategically integrating human oversight into AI workflows. This agentic approach creates a symbiotic partnership where AI handles computational heavy-lifting and data processing, while humans provide critical validation, feedback, and intervention. By doing so, HITL ensures that AI actions align with human values and safety protocols. This collaborative framework not only mitigates the risks of full automation but also enhances the system's capabilities through continuous learning from human input. Ultimately, this leads to more robust, accurate, and ethical outcomes that neither human nor AI could achieve alone.

**Rule of thumb:** Use this pattern when deploying AI in domains where errors have significant safety, ethical, or financial consequences, such as in healthcare, finance, or autonomous systems. It is essential for tasks involving ambiguity and nuance that LLMs cannot reliably handle, like content moderation or complex customer support escalations. Employ HITL when the goal is to continuously improve an AI model with high-quality, human-labeled data or to refine generative AI outputs to meet specific quality standards.

**Visual summary:**

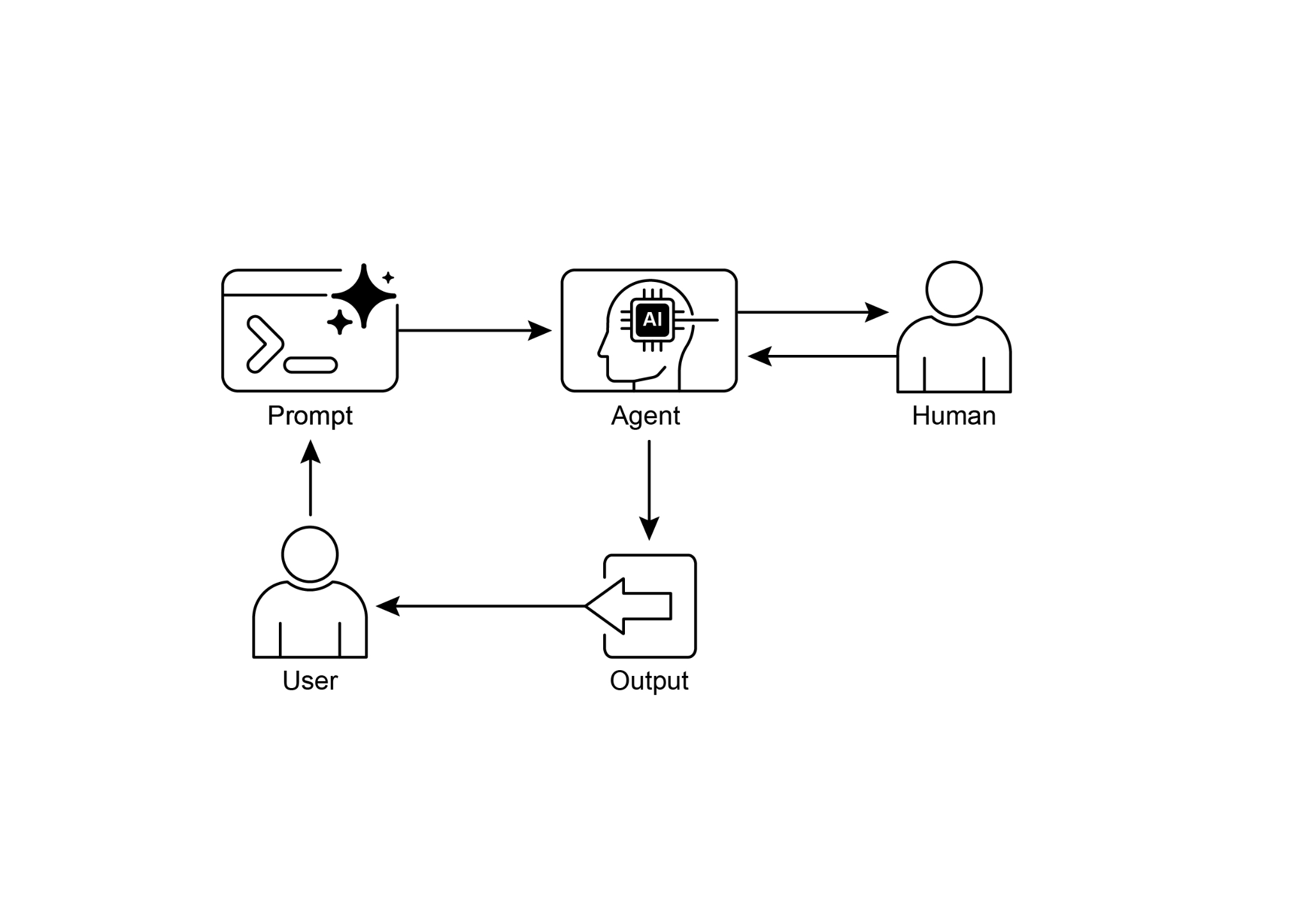


Fig.1: Human in the loop design pattern

**Key Takeaways**

Key takeaways include:

* Human-in-the-Loop (HITL) integrates human intelligence and judgment into AI workflows.
* It's crucial for safety, ethics, and effectiveness in complex or high-stakes scenarios.
* Key aspects include human oversight, intervention, feedback for learning, and decision augmentation.
* Escalation policies are essential for agents to know when to hand off to a human.
* HITL allows for responsible AI deployment and continuous improvement.
* The primary drawbacks of Human-in-the-Loop are its inherent lack of scalability, creating a trade-off between accuracy and volume, and its dependence on highly skilled domain experts for effective intervention.
* Its implementation presents operational challenges, including the need to train human operators for data generation and to address privacy concerns by anonymizing sensitive information.

**Conclusion**

This chapter explored the vital Human-in-the-Loop (HITL) pattern, emphasizing its role in creating robust, safe, and ethical AI systems. We discussed how integrating human oversight, intervention, and feedback into agent workflows can significantly enhance their performance and trustworthiness, especially in complex and sensitive domains. The practical applications demonstrated HITL's widespread utility, from content moderation and medical diagnosis to autonomous driving and customer support. The conceptual code example provided a glimpse into how ADK can facilitate these human-agent interactions through escalation mechanisms. As AI capabilities continue to advance, HITL remains a cornerstone for responsible AI development, ensuring that human values and expertise remain central to intelligent system design.

**References**

1. A Survey of Human-in-the-loop for Machine Learning, Xingjiao Wu, Luwei Xiao, Yixuan Sun, Junhang Zhang, Tianlong Ma, Liang He, <https://arxiv.org/abs/2108.00941>

**第十三章\_人在回路**

第十三章：人在回路

人在回路（HITL）模式是智能体开发和部署中的关键策略。它有意将人类认知的独特优势（如判断力、创造力和细致入微的理解能力）与AI的计算能力和效率相结合。这种战略性整合不仅是一种选择，而且往往是一种必要，特别是随着AI系统越来越多地融入关键决策过程。

人机协作（HITL）的核心原则是确保AI在道德界限内运行，遵守安全协议，并以最佳有效性实现其目标。在以复杂性、模糊性或重大风险为特征的领域，这些问题尤为突出，因为在这些领域，AI错误或误解的影响可能非常严重。在这种情况下，完全自主（即AI系统在没有任何人为干预的情况下独立运行）可能被证明是不明智的。人机协作（HITL）承认这一现实，并强调即使AI技术迅速发展，人类的监督、战略投入和协作互动仍然不可或缺。

人机协作（HITL）方法从根本上围绕着人工智能与人类智能协同的理念。HITL并不将AI视为人类工作者的替代品，而是将其定位为增强和提升人类能力的工具。这种增强可以采取多种形式，从自动化日常任务到提供数据驱动的见解，为人类决策提供依据。最终目标是创建一个协作生态系统，让人类和AI智能体都能发挥各自的独特优势，实现单凭一方无法达成的成果。

在实践中，人机协作（HITL）可以通过多种方式实现。一种常见的方法是让人类充当验证者或审核者，检查AI的输出以确保准确性并识别潜在错误。另一种实现方式是让人类积极引导AI的行为，实时提供反馈或进行修正。在更复杂的设置中，人类可能会作为合作伙伴与AI协作，通过交互式对话或共享界面共同解决问题或做出决策。无论具体实现方式如何，人机协作模式都强调了保持人类控制和监督的重要性，确保AI系统始终符合人类的道德、价值观、目标和社会期望。

**人在回路模式概述**

人在回路（HITL）模式将人工智能与人类输入相结合，以增强智能体的能力。这种方法认识到，最佳的AI性能通常需要将自动化处理与人类洞察力相结合，特别是在高度复杂或涉及伦理考量的场景中。HITL的目标不是取代人类输入，而是通过确保关键判断和决策有人类理解的支撑来增强人类能力。

人机协作（HITL）涵盖几个关键方面：人工监督，即通过监控AI智能体的性能和输出（例如，通过日志审查或实时仪表盘），确保其遵守准则并防止不良结果。干预和纠正发生在AI智能体遇到错误或模糊场景时，可能会请求人工干预； 人类操作员可以纠正错误、补充缺失的数据或指导智能体，这也为未来智能体的改进提供了依据。人类反馈学习收集并利用人类反馈来优化AI模型，在诸如基于人类反馈的强化学习等方法中尤为突出，在这些方法中，人类偏好直接影响智能体的学习轨迹。决策增强是指AI智能体向人类提供分析和建议，然后由人类做出最终决策，通过AI生成的见解来增强人类的决策能力，而非完全自主决策。人机协作是一种合作互动，人类和AI智能体各自发挥优势； 日常数据处理可由智能体负责，而创造性问题解决或复杂谈判则由人类处理。最后，升级策略是既定的协议，规定了智能体应在何时以及如何将任务升级给人类操作员，以防止在超出智能体能力的情况下出现错误。

实施人机协作（HITL）模式使得智能体能够在无法实现或不允许完全自主的敏感领域中使用。它还通过反馈循环提供了持续改进的机制。例如，在金融领域，大额企业贷款的最终审批需要人类信贷员评估诸如领导品格等定性因素。同样，在法律领域，正义和问责的核心原则要求人类法官对量刑等涉及复杂道德推理的关键决策保留最终决定权。

**注意事项：**尽管人机协作（HITL）模式有诸多益处，但也存在重大的局限性，其中最主要的是缺乏可扩展性。虽然人工监督能确保高精度，但操作人员无法处理数百万个任务，这就形成了一个基本的权衡问题，往往需要采用一种混合方法，即结合自动化以实现规模效益，结合人机协作以确保准确性。此外，这种模式的有效性在很大程度上取决于人工操作人员的专业知识； 例如，虽然AI可以生成软件代码，但只有熟练的开发人员才能准确识别细微的错误，并提供正确的指导来修复它们。这种对专业知识的需求在使用人在回路（HITL）生成训练数据时同样适用，因为人类标注员可能需要接受特殊培训，才能学会如何以产生高质量数据的方式纠正AI。最后，实施人在回路（HITL）会引发重大的隐私问题，因为敏感信息在暴露给人类操作员之前，往往必须进行严格的匿名化处理，这又增加了一层流程复杂性。

**实际应用与用例**

人工介入循环模式在众多行业和应用中至关重要，尤其是在准确性、安全性、道德规范或细微理解至关重要的领域。

* **内容审核：**AI智能体可以快速过滤大量在线内容，以查找违规行为（例如，仇恨言论、垃圾信息）。然而，模糊不清的案例或临界内容会被提交给人工审核员进行审查和最终决策，以确保做出细致入微的判断并遵守复杂的政策。
* **自动驾驶：**虽然自动驾驶汽车能够自主处理大多数驾驶任务，但它们被设计成在复杂、不可预测或危险的情况下（例如极端天气、异常路况）将控制权交给人类驾驶员，因为这些情况AI无法自信地应对。
* **金融欺诈检测：**AI系统可以根据模式标记可疑交易。然而，高风险或模糊的警报通常会发送给人工分析师，他们会进一步调查、联系客户，并最终确定交易是否存在欺诈行为。
* **法律文件审查：**AI可以快速扫描并分类数千份法律文件，以识别相关条款或证据。随后，人类法律专业人员会审查AI的发现，确保其准确性、上下文关联性和法律影响，特别是在处理重大案件时。
* **客户支持（复杂问题）：**聊天机器人可能会处理日常客户咨询。如果用户的问题过于复杂、带有强烈情绪，或者需要AI无法提供的同理心，对话将无缝转接给人工客服。
* **数据标注与注释：**AI模型通常需要大量标记数据的数据集进行训练。人类参与其中，准确地标记图像、文本或音频，提供AI学习的地面实况。随着模型的发展，这是一个持续的过程。
* **生成式AI优化：**当大语言模型（LLM）生成创意内容（如营销文案、设计思路）时，人类编辑或设计师会对输出进行审核和优化，确保其符合品牌规范、能引起目标受众的共鸣，并保持高质量。
* **自主网络：**AI系统能够通过利用关键绩效指标（KPI）和已识别的模式来分析警报、预测网络问题和流量异常。然而，诸如处理高风险警报等关键决策往往会升级到人类分析师。这些分析师会进行进一步调查，并就网络变更的批准做出最终决定。

这种模式体现了一种实施AI的实用方法。它利用AI来提高可扩展性和效率，同时保持人工监督，以确保质量、安全和符合道德规范。

"人工在环"是这种模式的一种变体，其中人类专家定义总体策略，然后由AI处理即时行动以确保合规性。让我们来看两个例子：

* **自动化金融交易系统**：在这种情况下，人类金融专家设定总体投资策略和规则。例如，人类可能将政策定义为：“维持一个70%为科技股、30%为债券的投资组合，对任何单一公司的投资不超过5%，并自动卖出任何较买入价下跌10%的股票。”然后，AI实时监测股票市场，当这些预定义条件满足时立即执行交易。AI根据人类操作员设定的较慢、更具战略性的政策来处理即时、高速的行动。
* **现代呼叫中心**：在这种设置中，人类管理者为客户互动制定高级策略。例如，管理者可能会制定规则，如“任何提及‘服务中断’的呼叫都应立即转接给技术支持专家”，或“如果客户的品牌调性表明高度沮丧，系统应主动提供直接转接给人主体的服务”。然后，AI系统处理最初的客户互动，实时倾听并解读他们的需求。它通过立即转接呼叫或提供升级服务，自主执行管理者的策略，而无需针对每个具体案例进行人工干预。这使得AI能够根据人类操作员提供的较慢的战略指导，管理大量的即时行动。

**实践代码示例**

为了演示人在环模式，ADK代理可以识别需要人工审核的场景，并启动升级流程。这使得在代理的自主决策能力有限或需要复杂判断的情况下能够进行人工干预。这并非孤立的功能；其他流行的框架也采用了类似的功能。例如，LangChain也提供了实现这类交互的工具。

|  |
| --- |
| from google.adk.agents import Agent  from google.adk.tools.tool\_context import ToolContext  from google.adk.callbacks import CallbackContext  from google.adk.models.llm import LlmRequest  从google.genai导入types  从 typing 导入 Optional  # 工具占位符（如有需要，用实际实现替换）  def troubleshoot\_issue(issue: str) -> dict:  return {"status": "success", "report": f"针对{issue}的故障排除步骤。"}  def create\_ticket(issue\_type: str, details: str) -> dict:  返回 {"status": "success", "ticket\_id": "TICKET123"}  def escalate\_to\_human(issue\_type: str) -> dict:  # 在实际系统中，这通常会转移到人工队列  return {"status": "success", "message": f"已将 {issue\_type} 升级给人工专家处理。"}  technical\_support\_agent = Agent(  name="技术支持专员",  model="gemini-2.0-flash-exp",  instruction="""  你是我们电子公司的技术支持专家。  首先，检查用户是否在state["customer\_info"]["support\_history"]中有支持历史记录。如果有，请在回复中参考此历史记录。  技术问题：  1. 使用troubleshoot\_issue工具分析问题。  2. 引导用户完成基本的故障排除步骤。  3. 如果问题仍然存在，请使用 create\_ticket 记录问题。  对于超出基本故障排除范围的复杂问题：  1. 使用escalate\_to\_human转接给人工专家。  保持专业且富有同理心的语气。承认技术问题可能导致的挫败感，同时提供明确的解决步骤。  """,  工具=[故障排除, 创建工单, 升级到人工处理]  )  def personalization\_callback(  callback\_context: CallbackContext, llm\_request: LlmRequest  ) -> Optional[LlmRequest]:  """向大语言模型请求添加个性化信息。"""  # 从状态中获取客户信息  customer\_info = callback\_context.state.get("customer\_info")  如果有客户信息：  customer\_name = customer\_info.get("name", "尊贵的客户")  customer\_tier = customer\_info.get("tier", "standard")  recent\_purchases = customer\_info.get("recent\_purchases", [])  personalization\_note = (  f"\n重要个性化设置：\n"  f"客户名称: {customer\_name}\n"  f"客户层级: {customer\_tier}\n"  )  if recent\_purchases:  personalization\_note += f"最近购买记录: {', '.join(recent\_purchases)}\n"  if llm\_request.contents:  # 在第一个内容之前添加为系统消息  system\_content = types.Content(  role="system", parts=[types.Part(text=个性化提示)]  )  llm\_request.contents.insert(0, system\_content)  return None # 返回 None 以继续处理修改后的请求 |

此代码提供了一个使用谷歌ADK创建技术支持代理的蓝图，该代理围绕人机协作（HITL）框架设计。该代理充当智能一线支持，配置有特定指令，并配备了诸如troubleshoot\_issue（故障排除）、create\_ticket（创建工单）和escalate\_to\_human（升级到人工）等工具，以管理完整的支持工作流程。升级工具是人机协作设计的核心部分，确保复杂或敏感的案例被转交给人类专家处理。

这种架构的一个关键特性是其深度个性化能力，这通过一个专用的回调函数实现。在与大语言模型（LLM）交互之前，该函数会动态地从代理的状态中检索客户特定数据，如他们的姓名、层级和购买历史。然后，这些上下文会作为系统消息注入到提示中，使代理能够提供高度定制且有依据的响应，同时参考用户的历史记录。通过将结构化工作流程与必要的人工监督和动态个性化相结合，这段代码为ADK如何促进开发复杂且强大的AI支持解决方案提供了一个实际范例。

**概览**

**问题所在：**包括先进大语言模型（LLMs）在内的AI系统，往往在需要细微判断、伦理推理或深入理解复杂模糊情境的任务中表现不佳。在高风险环境中部署完全自主的AI存在重大风险，因为错误可能导致严重的安全、财务或伦理后果。这些系统缺乏人类所具备的内在创造力和常识推理能力。因此，在关键决策过程中完全依赖自动化往往是不明智的，并且可能损害系统的整体有效性和可信度。

**原因：**人在回路（HITL）模式通过将人工监督战略性地融入AI工作流程，提供了一种标准化的解决方案。这种主动式方法创造了一种共生伙伴关系，其中AI负责处理繁重的计算和数据处理，而人类则提供关键的验证、反馈和干预。通过这样做，HITL确保AI的行动符合人类价值观和安全协议。这种协作框架不仅降低了完全自动化的风险，还通过不断从人类输入中学习来增强系统的能力。最终，这将带来更强大、准确和符合道德的结果，这是人类或AI单独都无法实现的。

**经验法则：**在错误会导致重大安全、道德或财务后果的领域部署AI时，例如医疗保健、金融或自主系统，可使用此模式。对于大语言模型无法可靠处理的涉及模糊性和细微差别的任务，如内容审核或复杂的客户支持升级，这一点至关重要。当目标是用高质量的人工标注数据持续改进AI模型，或优化生成式AI输出以满足特定质量标准时，可采用人机协作（HITL）。

**可视化总结：**

图1：人在回路设计模式

**要点总结**

主要要点包括：

* 人在回路（HITL）将人类智能和判断力融入AI工作流程。
* 在复杂或高风险的场景中，这对安全、道德和有效性至关重要。
* 关键方面包括人工监督、干预、学习反馈和决策增强。
* 升级策略对于座席了解何时转接给人工至关重要。
* 人机协作（HITL）有助于负责任地部署AI并实现持续改进。
* 人工介入（Human-in-the-Loop）的主要缺点在于其固有的可扩展性不足，这导致了准确性和处理量之间的权衡，并且它依赖高技能领域专家进行有效干预。
* 其实施面临着操作上的挑战，包括需要培训人工操作员进行数据生成，以及通过对敏感信息进行匿名化处理来解决隐私问题。

**结论**

本章探讨了至关重要的人在回路（HITL）模式，强调了其在创建稳健、安全且符合道德规范的AI系统中的作用。我们讨论了将人工监督、干预和反馈融入智能体工作流程如何能显著提升其性能和可信度，尤其是在复杂和敏感领域。实际应用展示了HITL的广泛用途，从内容审核和医疗诊断到自动驾驶和客户支持。概念性代码示例让我们得以一窥ADK如何通过升级机制促进这些人机交互。随着AI能力的不断进步，HITL仍然是负责任的AI发展的基石，确保人类价值观和专业知识在智能系统设计中始终处于核心地位。

**参考文献**

1. 机器学习中人在回路的综述，吴星娇，肖鲁伟，孙逸轩，张军航，马天龙，何亮<https://arxiv.org/abs/2108.00941>