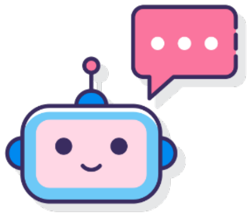
COMP 3647
Human-AI Interaction Design

Topic 10
Human-in-the-Loop in Trustworthy Autonomous Systems (TAS)

Prof. Effie L-C Law

Chatbots: Short Chitchat



- Human-in-the-Loop?
- It's all about Trust.
- Really?
- Hmm….
- Trust what?
- I'm talking about Autonomous Systems
- Okay, are they trustworthy?
- Yes. Maybe. Not sure
- Can I trust you?
- Sure, why not?

Overview

Concepts and Applications of Human-in-the-Loop (HiL)

- HiL in Human-Robot Team
- HiL in Machine Learning
- HiL in Autonomous Vehicle

Definition: Autonomous Systems (AS)

Autonomous systems

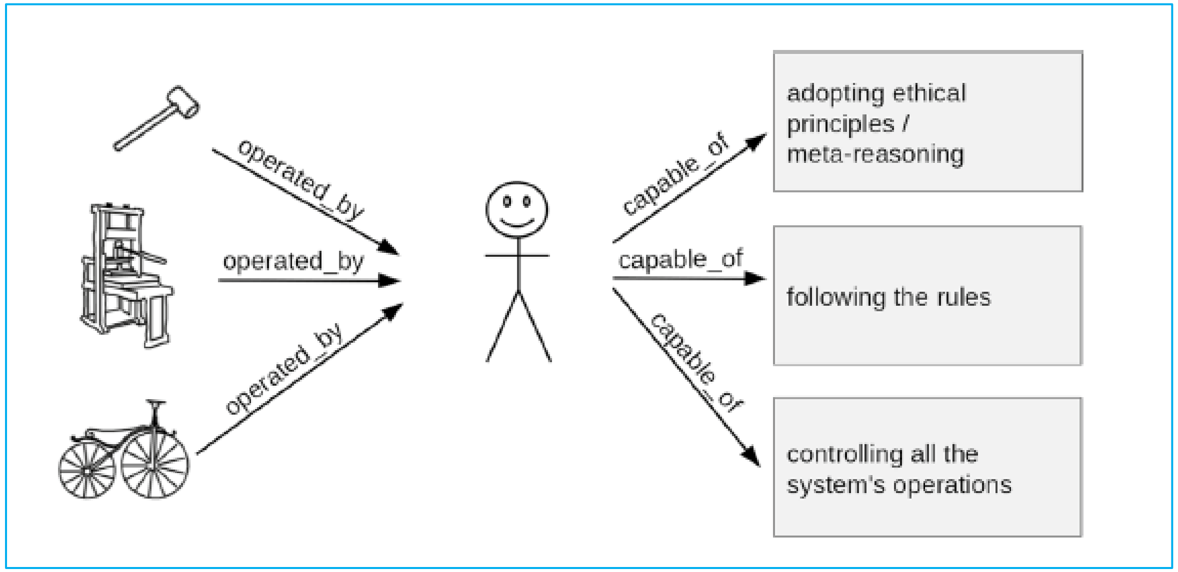
decide for themselves what to do and when to do it … varying in the degree of autonomy used, from almost **pure human control** to fully autonomous activities with **minimal human interaction**

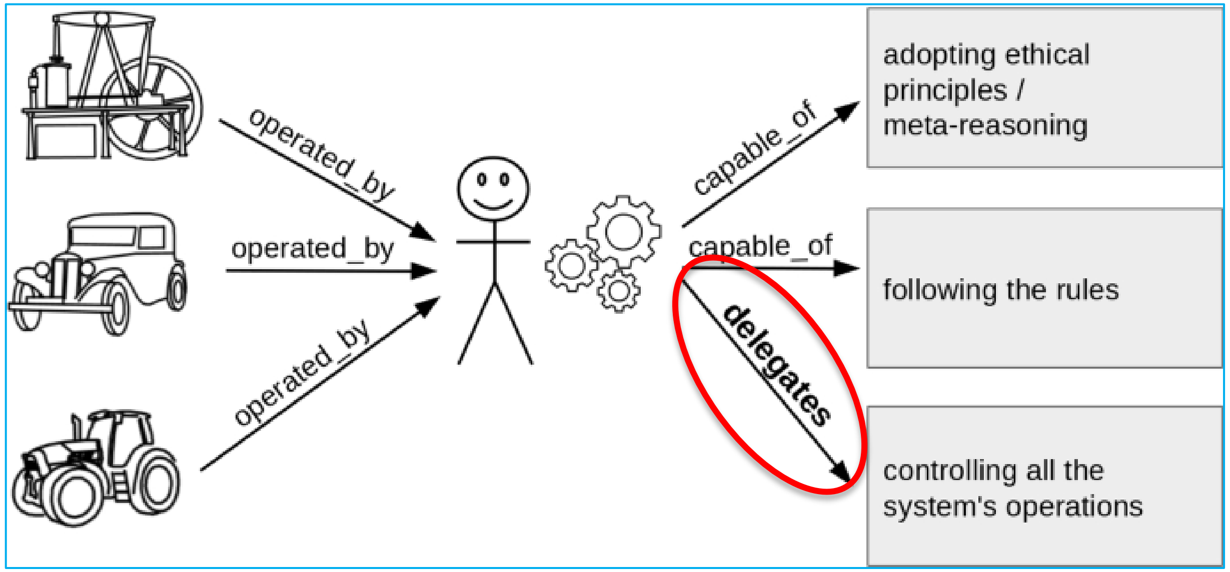
(Fisher et al. CACM, 2013)

Autonomy:

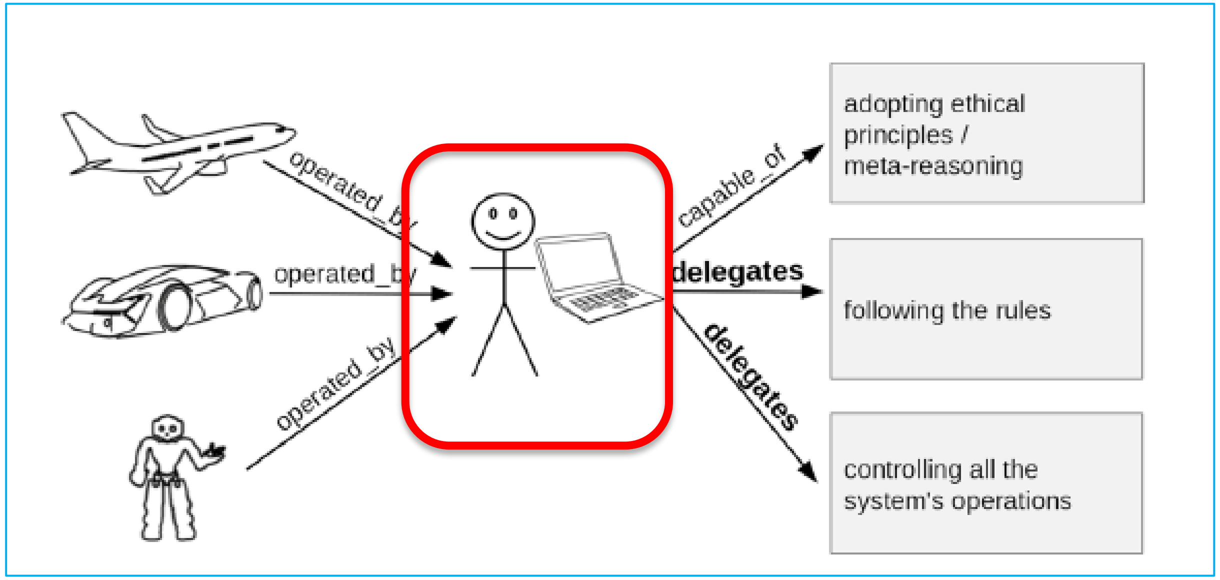
selective inclusion of human involvement in a task

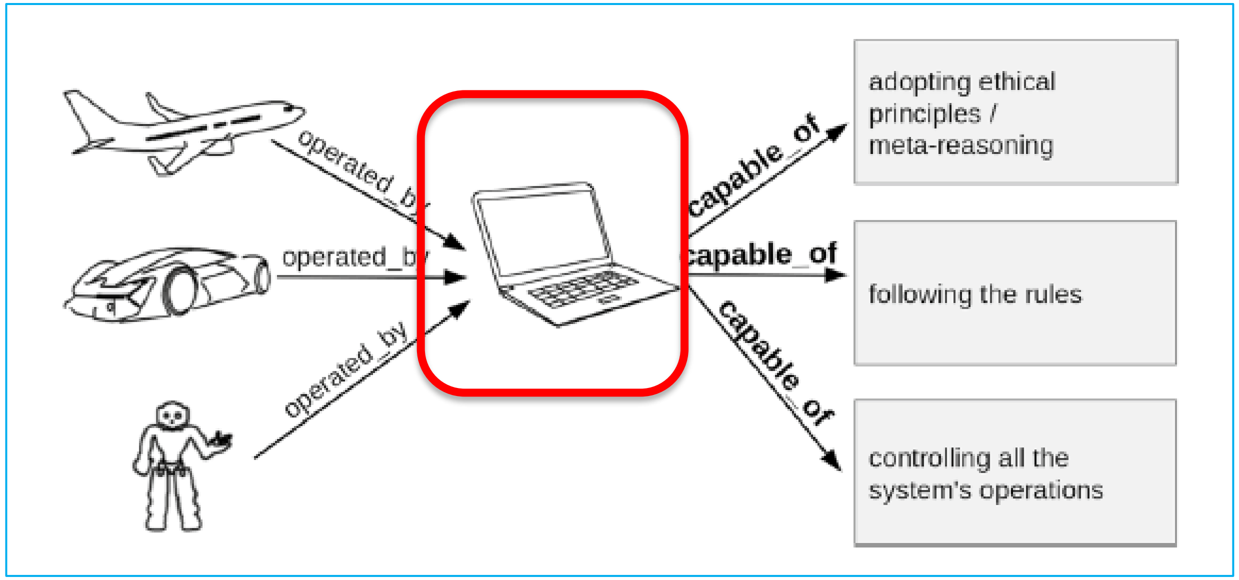
Autonomous Systems (Fisher et al. 2013)





Autonomous Systems (Fisher et al. 2013)





Autonomous Systems Principles

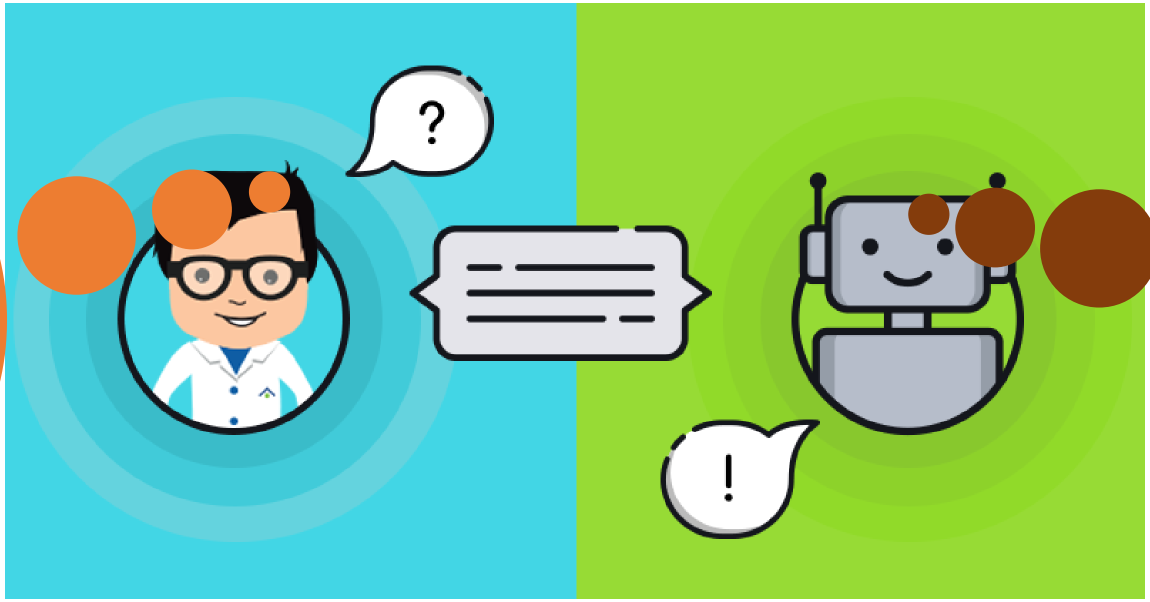
Wirsing et al. (2011) provide four major autonomous systems principles:

- **Knowledge:** The system knows facts about itself and its surroundings.
- **Adaptation:** The system can adapt its own behaviour dynamically to cope with changing surroundings.
- **Self-awareness:** The system can examine and reason about its own state.
- **Emergence:** Simple system elements construct complex entities.

AI’s User Model vs. User’s Mental Model

❖ A ***playful,*** social agent with personality

❖ A ***functional*** tool; Google in conversation



Human-human interaction:
- **Instrumentality**
- **Curiosity/Fun** *(“Tell me a joke!”)*

Natural Language Classifier:
- **Intent correctness**
- **Confidence level** (Low: *“Sorry I don’t know”*)

❑ A ***task-focused*** user expecting curated information

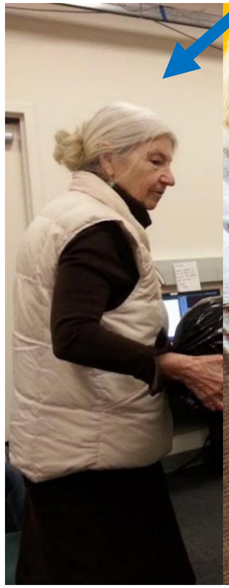
❑ A ***fun-seeking*** user expecting humanized responses

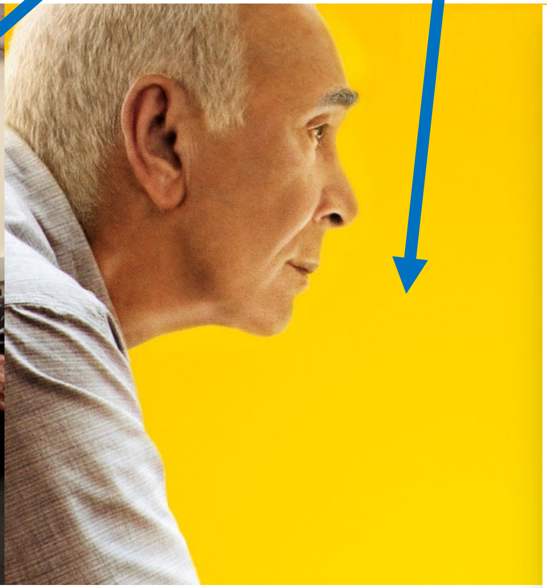
Human-in-the-Loop (HiL) in AS, Why?

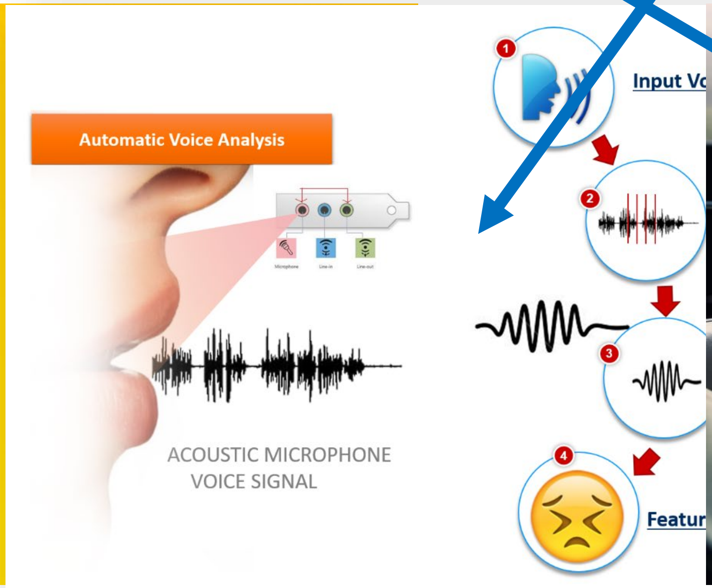
- Ensure high usability and positive user experience (UX) of AS, addressing human needs and preferences
- Handle complex tasks in unstructured/uncertain environments, combining cognitive skills of humans with AS behaviours
- Maximize accuracy, addressing incorrect predictions due to faulty AS's user models
- Elicit value-added content and features
- Meet safety, legal and financial regulations

Enhancing *Trustworthiness* (adding the *T* to AS)

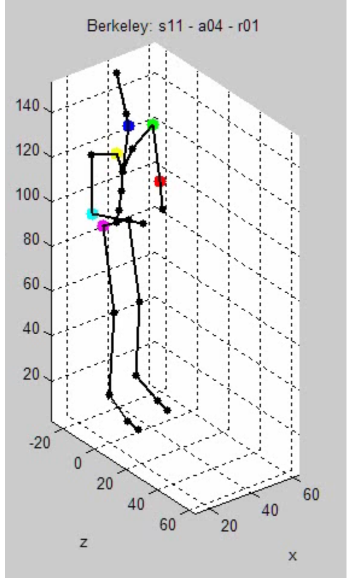
		Temporality of Human-AS Interaction	
		Synchronous	Asynchronous
Human Role	Participatory	**Human-Robot Teams** Collaborative assembly; Social robots	**Machine Learning** labelling, tagging, annotation
	Supervisory	**Highly/Fully Automated Systems** (talk-back-control)	Feedback on recorded robot performance

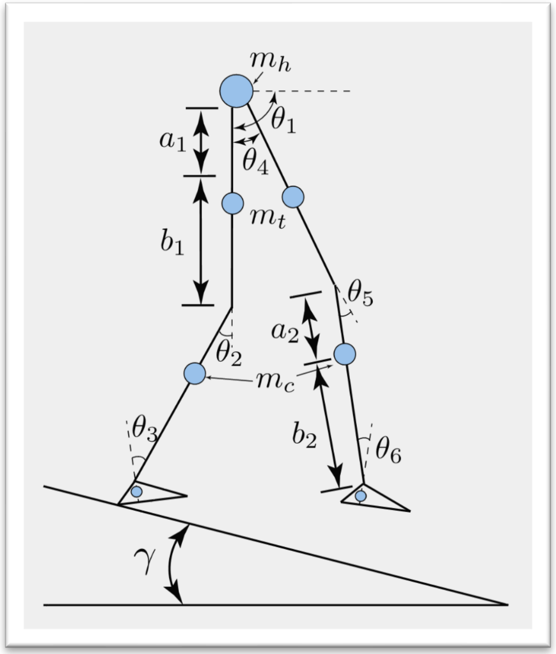






Human Models: Kinematics/Dynamics
(Ruzena Bajcsy, UC Berkeley)





Aim:
Design algorithms for optimizing human-robot control sharing

Fact:
Human is a complex kinematic/dynamic system with many degrees of freedom and parameters, which vary with individuals and activities.

Challenge:
Identify the appropriate representation of human physical actions for a specific application.

Human-Robot Collaborative Manipulation

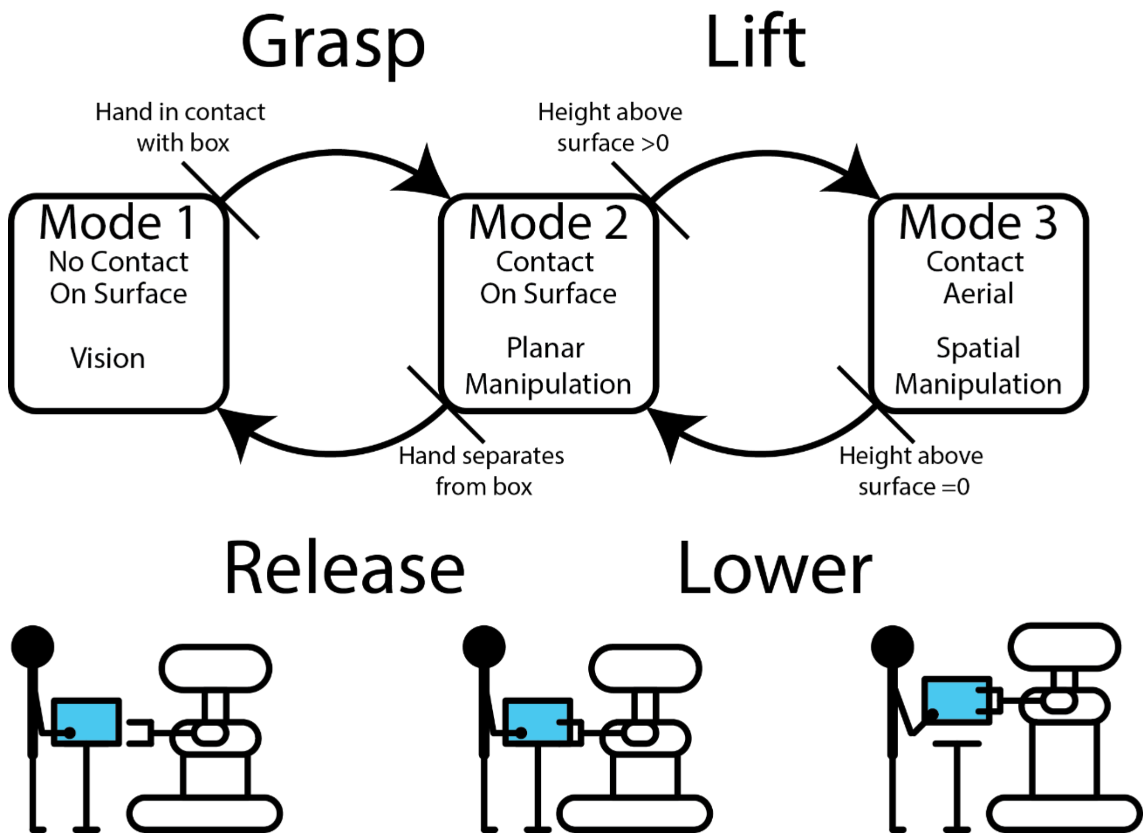
Ruzena Bajcsy, 2014

Aim:

Enable intelligent control of robots, providing direct physical assistance to humans.

Challenges:

- Create **unified** model of the human-robot coupled mechanical system
- Predict **intent** of human operator based on physical cues
- **Individualized**, data-driven modeling of human kinematics and dynamics

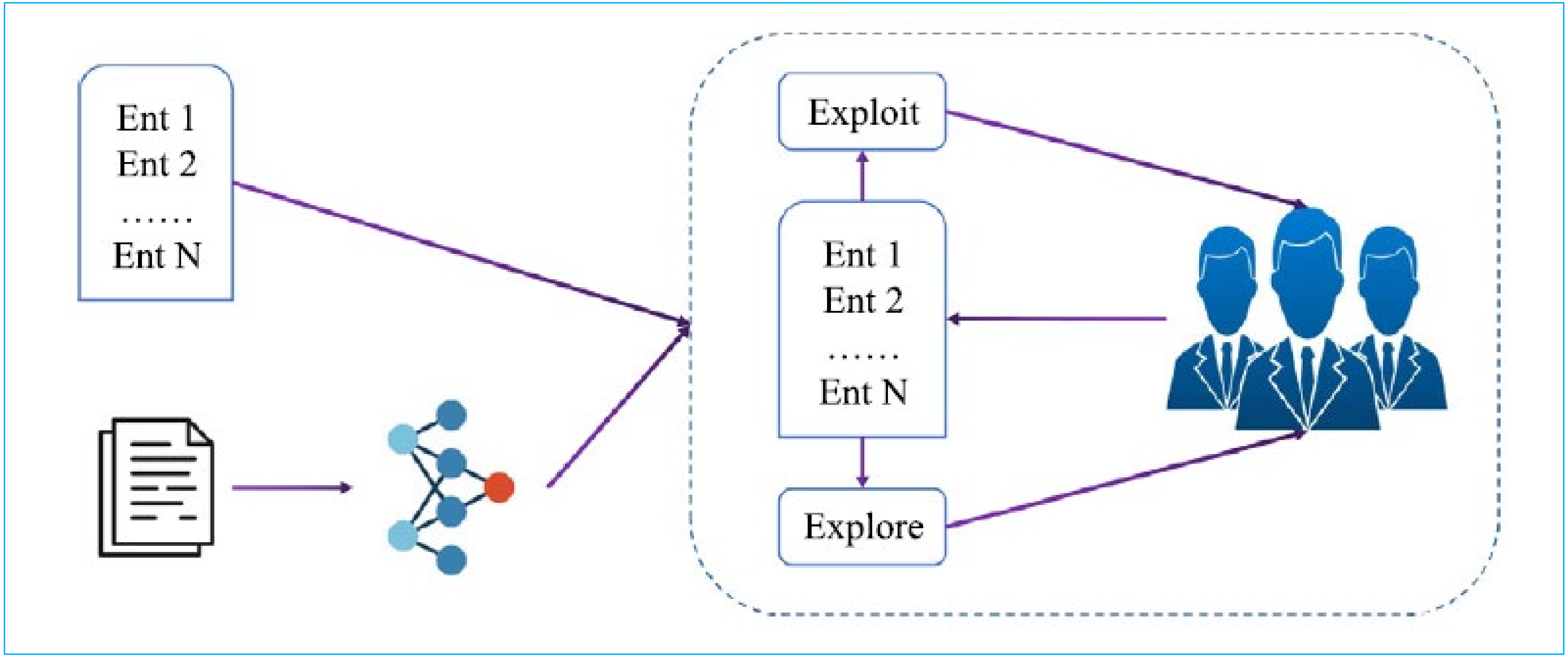


Robot-Assisted Dressing (RAD) – Healthcare & Assisted Technology (Mosuavi et al. 2023)

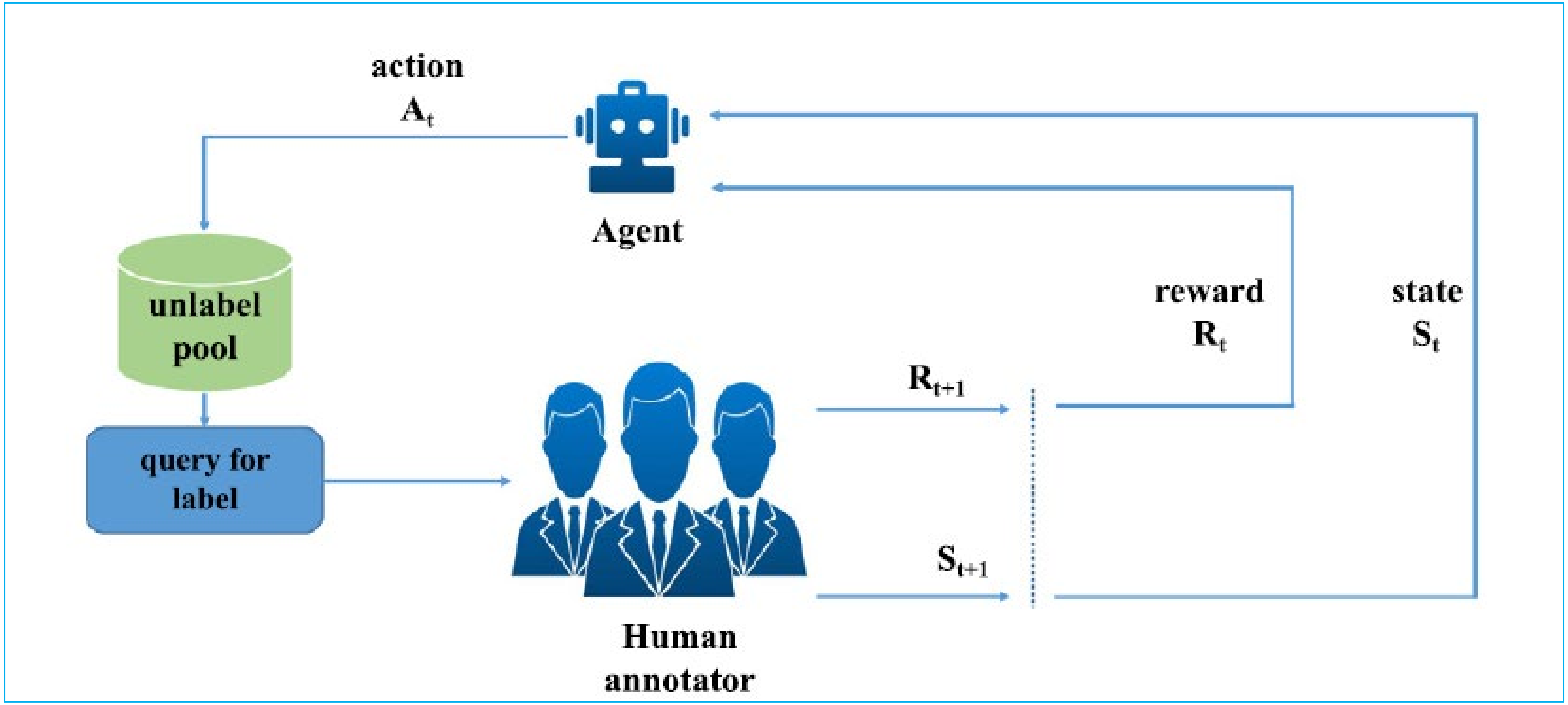
New Sawyer collaborative robot designed for precise, repetitive tasks in electronics assembly and testing (Courtesy of Rethink Robotics Inc.)

		Temporality of Human-Robot Transactions	
		Synchronous	Asynchronous
Human Role	Participatory (Active)	**Human-Robot Teams** Collaborative assembly; Social robots; Chatbots	**Machine Learning** labelling, tagging, annotation
	Supervisory (Passive)	**Highly/Fully Automated Systems** (talk-back-control)	Feedback on recorded robot performance

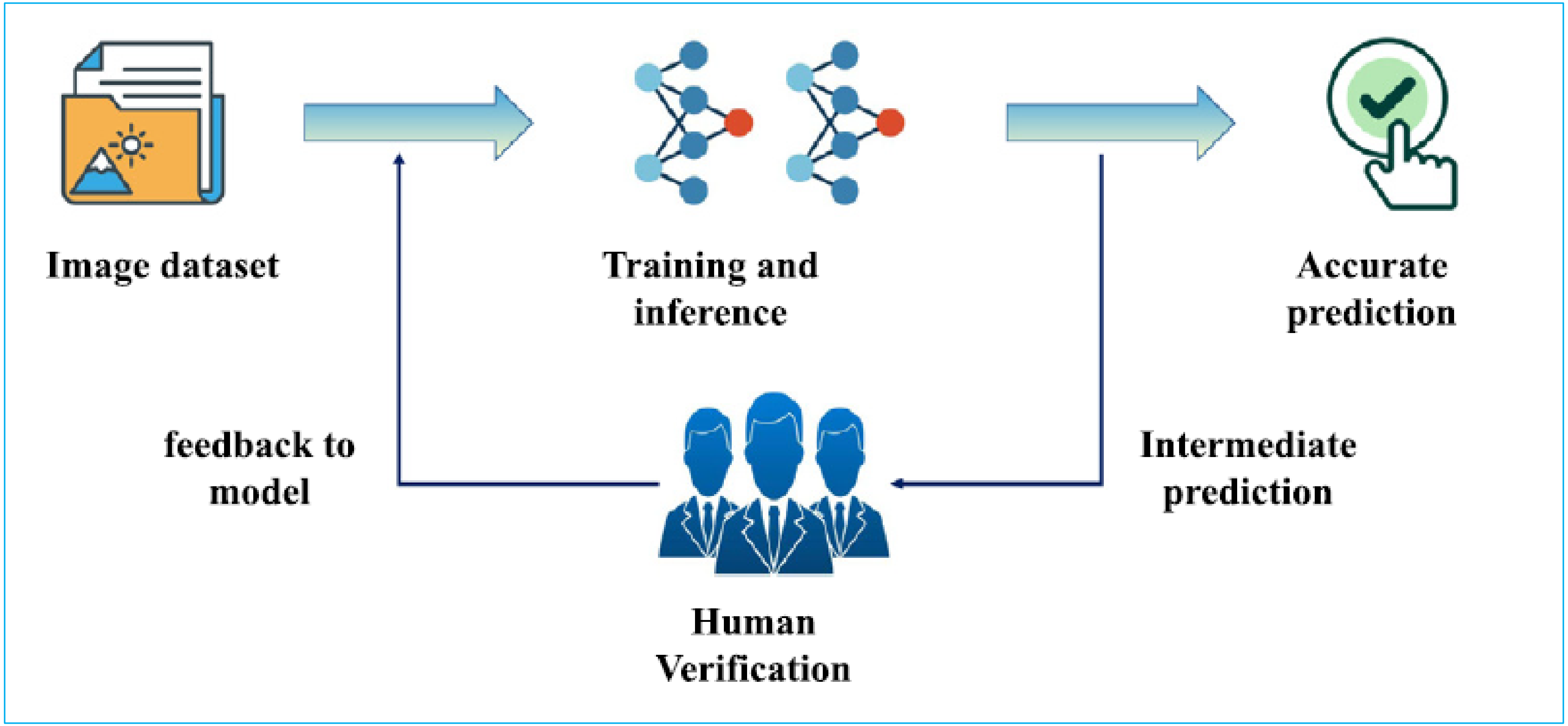
Human-in-the-Loop (Wu et al. 2020)



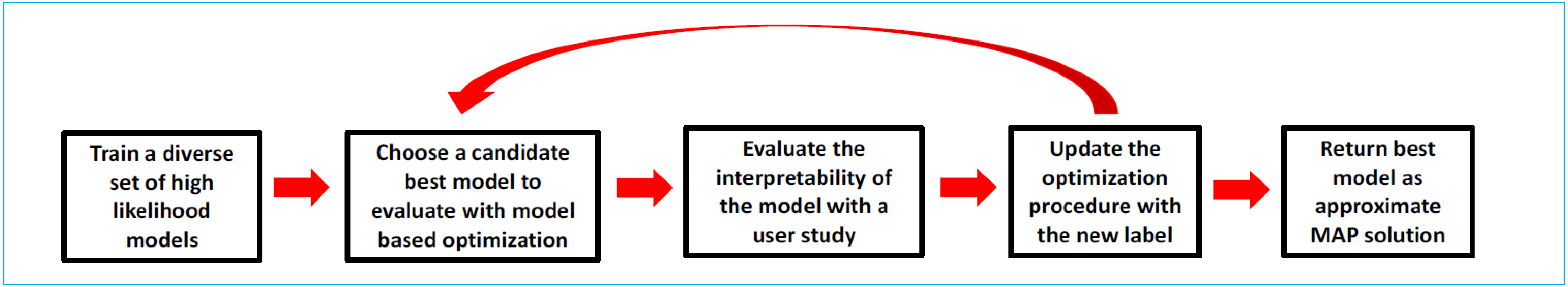
Human-in-the-Loop (Wu et al. 2020)



Human-in-the-Loop (Wu et al. 2020)



Interpretability: Human-in-the-Loop (Lage et al. 2018)



Interpretability vs. Explainability in AI

Explainability:

- **Definition:** Explainability refers to the ability to describe or provide reasons for how a model or system makes specific decisions or predictions.
- **Focus:** It emphasizes the communication of the model's decision-making process in a way that is understandable to end-users or stakeholders.
- **Example:** If a machine learning model predicts that a loan application is denied, an explanation would involve clarifying which features contributed to this decision, such as low credit score or high debt-to-income ratio.

Interpretability:

- **Definition:** Interpretability is the degree to which a human can understand the cause-and-effect relationships in a system. In the context of machine learning, it often refers to understanding the relationships between input features and the model's output.
- **Focus:** It is more concerned with the comprehensibility of the model itself, examining how changes in input variables affect the model's predictions.
- **Example:** In a linear regression model, interpretability would involve understanding how changes in each input feature linearly contribute to changes in the predicted output.

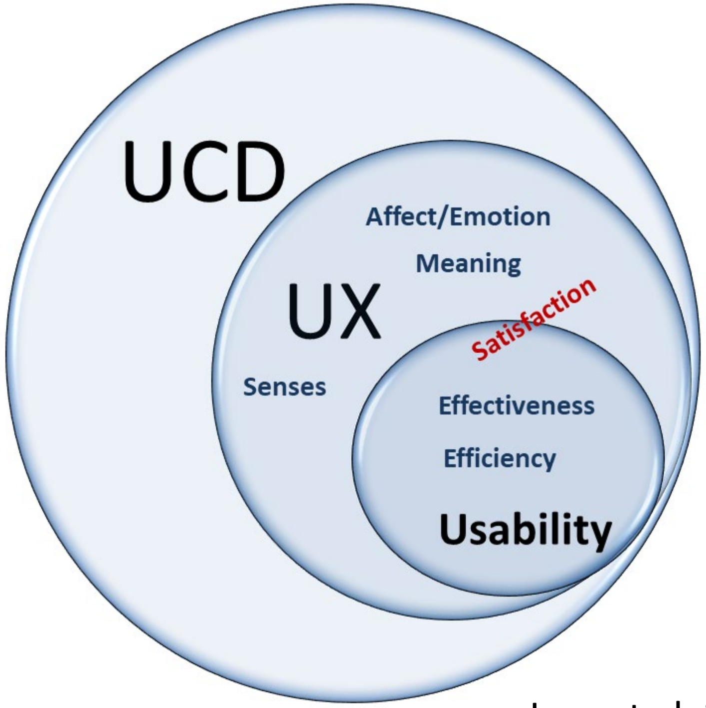
Durham University

		Temporality of Human-Robot Transactions	
		Synchronous	**Asynchronous**
Human Role	Participatory (Active)	**Human-Robot Teams** Collaborative assembly; Social robots;	**Machine Learning** labelling, tagging, annotation
	Supervisory (Passive)	**Highly/Fully Automated Systems** Sensor-based (talk-back-control)	Feedback on recorded robot performance

HiL in Cyber Physical Systems (HiL-CPS)

Interaction Design (IxD)

- Usability and User Experience (UX)
- Iterative prototyping (User-centred Design, UCD)
- Human factors: *attention, fatigue, stress*
- Trust: Interpretability, Explainability



Law et al. 2015

Control-sharing Strategies (Gil et al. 2019/20/22)

- Attention management with context-awareness
- Natural and understandable collaboration
- Optimal level of obtrusiveness for feedback

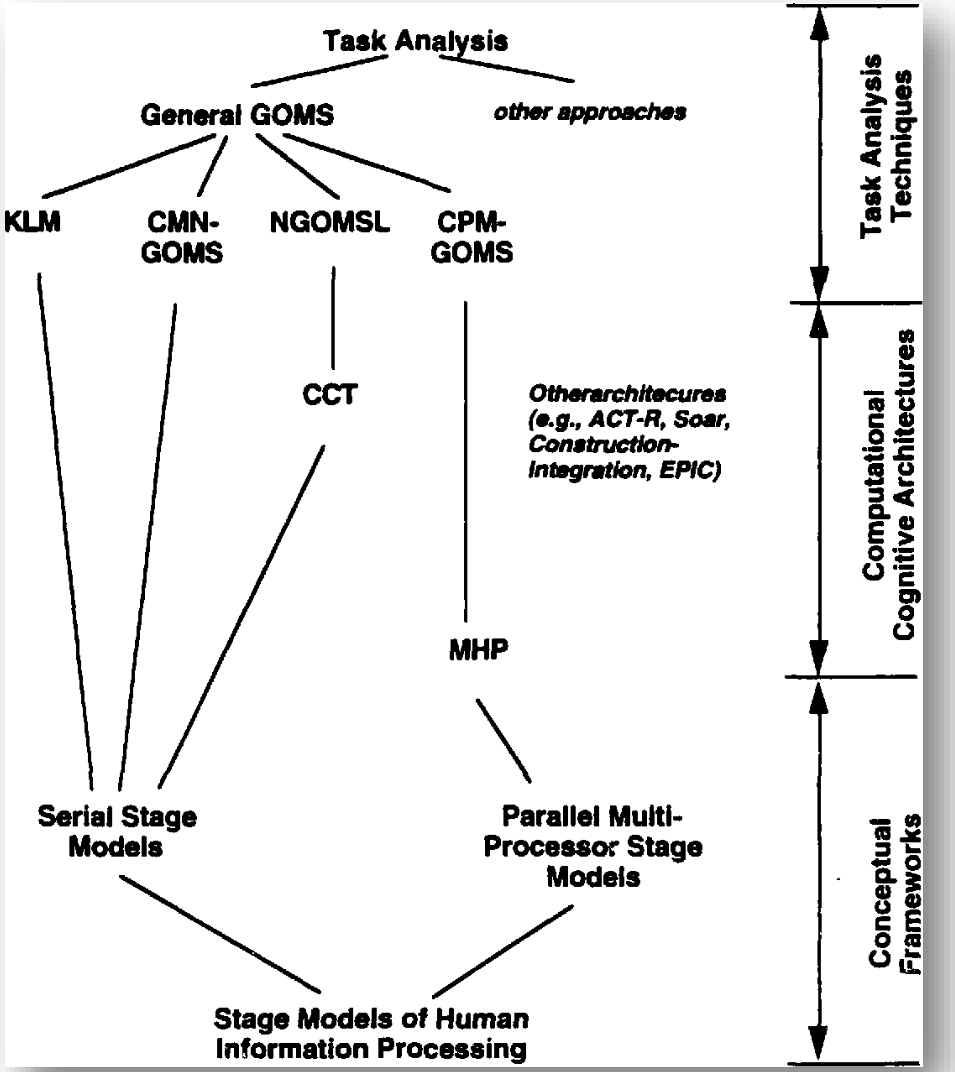
Challenges

- Human-computer integration → multisensory signals → mental states
- Human intent inferences → anticipate user intention and adapt

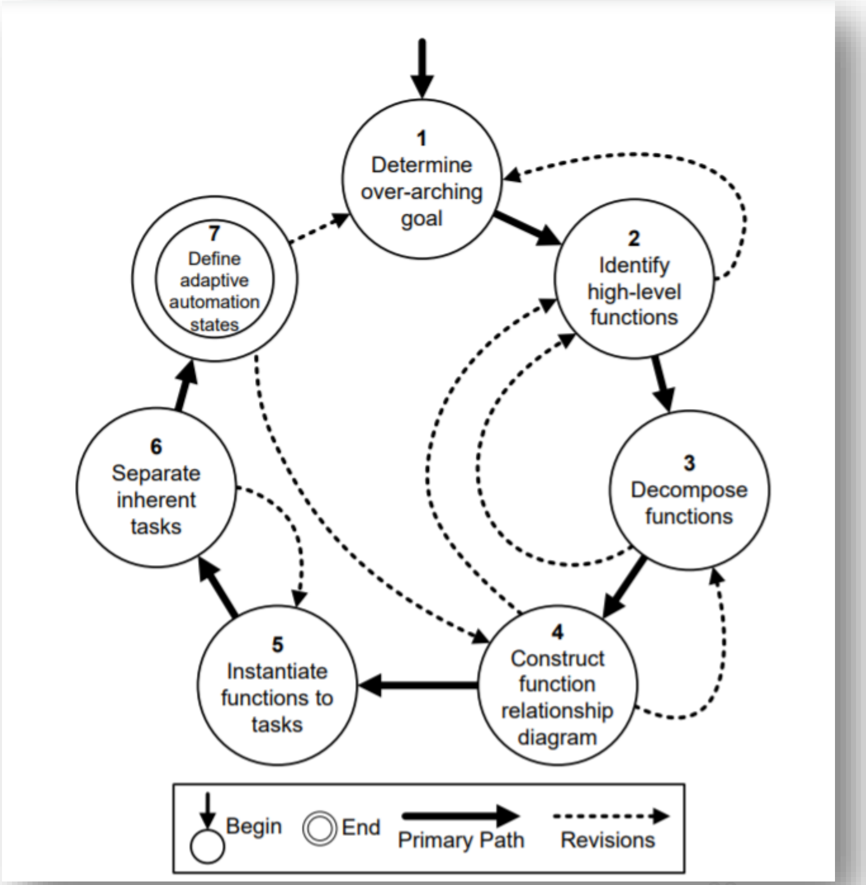
HiL-CPS: Function and Task Modelling

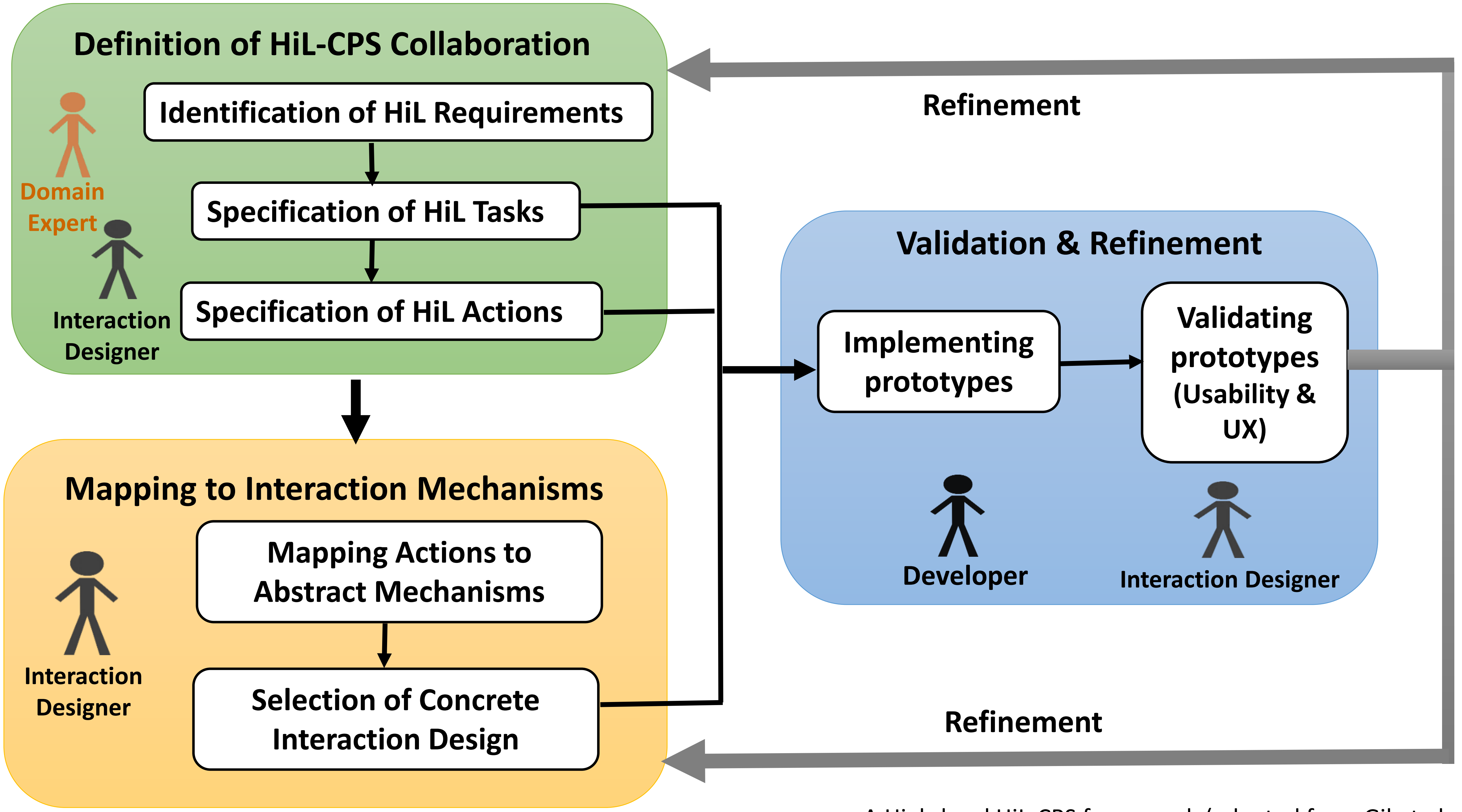
Task Analysis Models:

- GOMS (Goals, Operators, Methods, Selection Rules)
- HTA (Hierarchical Task Analysis)
- CTT(ConcurTree)



Function-to-Task Design Process Model
(Bindewald et al. 2014)





A High-level HiL-CPS framework (adapted from Gil et al., 2019)

HiL-CPS: Autonomous Vehicles (AV)

❖ **Takeover (Supervised Autonomous Driving)**

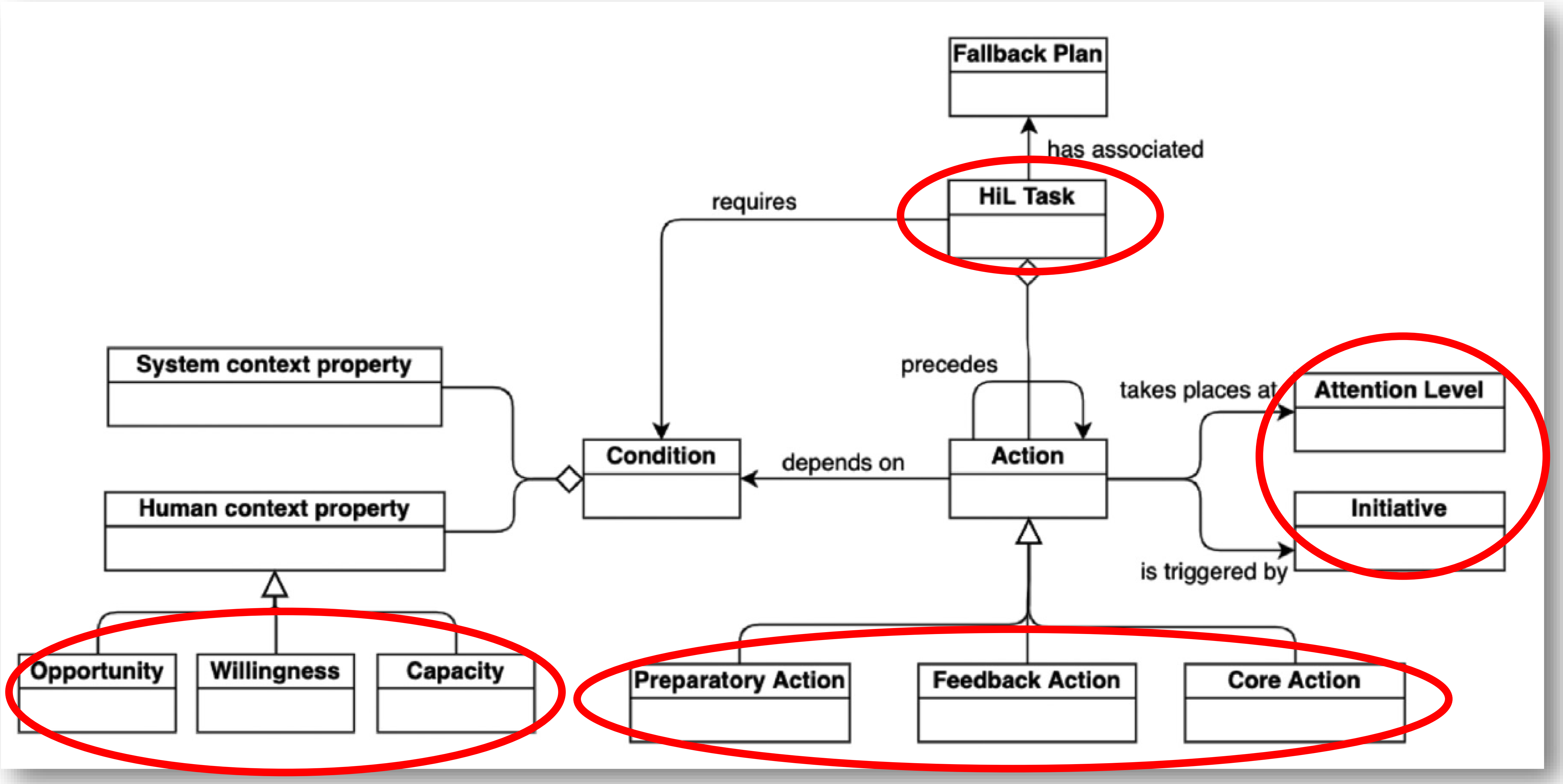
Car ------- CONTROL ------> Human

❖ **Handover (Supervised Manual Driving)**

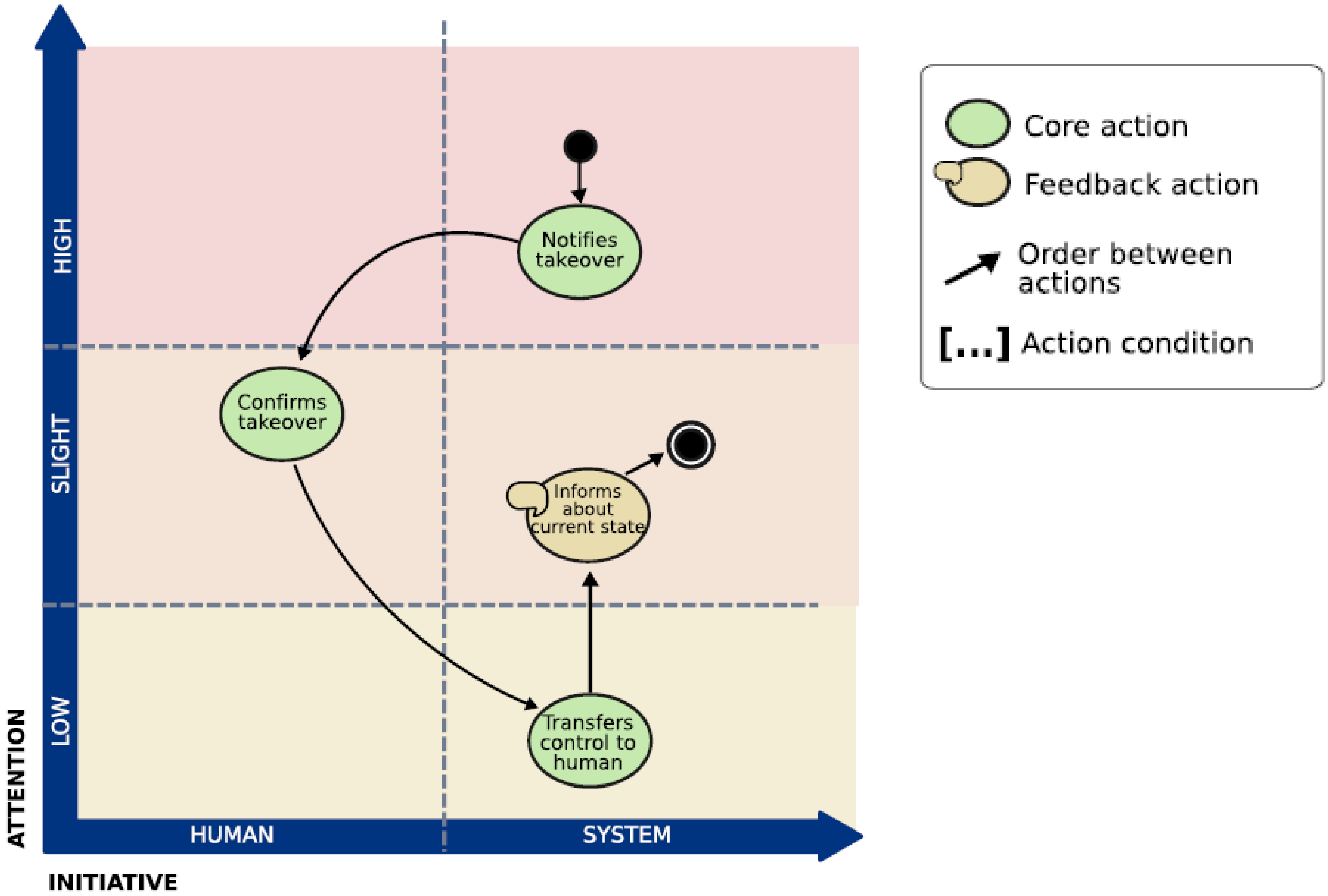
Human -------- CONTROL ------> Car

Types of control transfer:

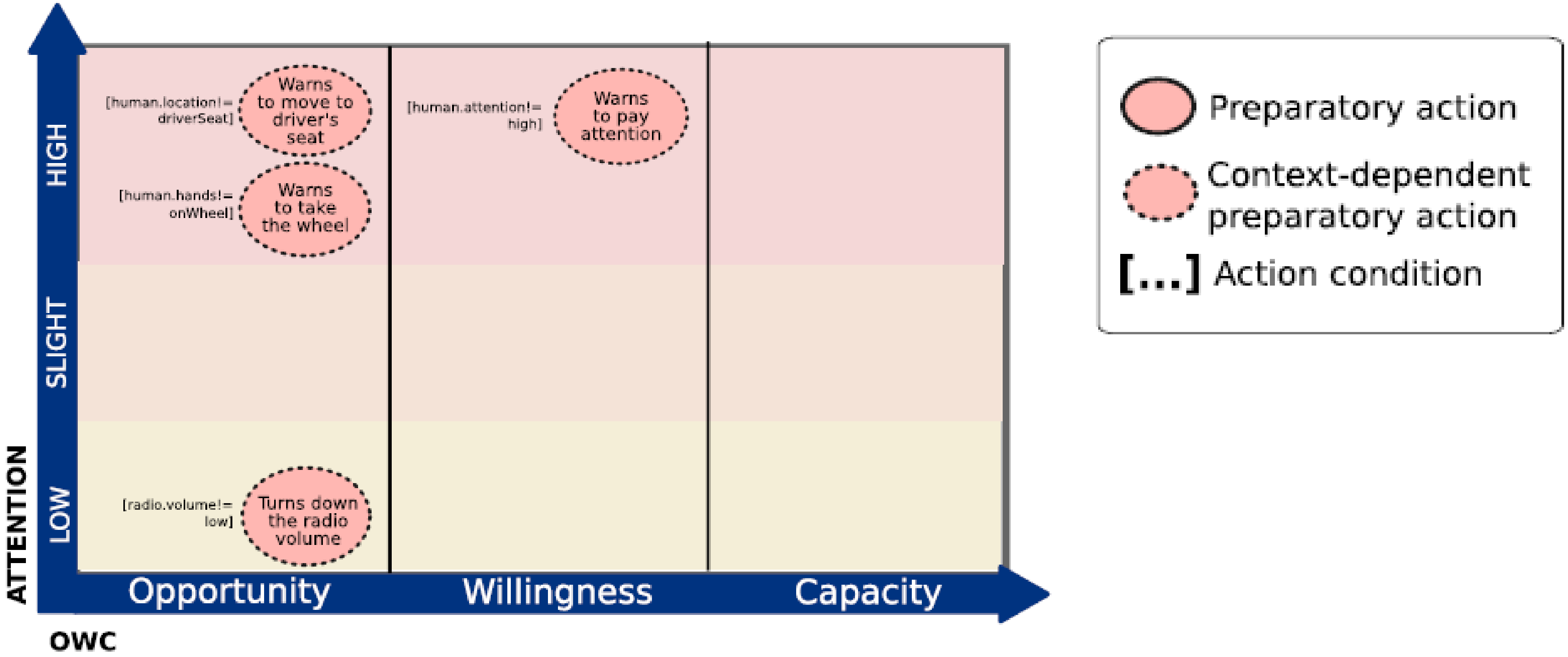
- Anticipated Takeover (e.g. entering a busy city centre)
- Emergency Takeover (e.g. near crash scenario)
- Anticipated Handover (e.g. receiving a scheduled phone call)
- Emergency Handover (e.g. sleepy/fatigue driver)



Gil et al. 2020



HIGH
SLIGHT
LOW
ATTENTION
HUMAN
SYSTEM
INITIATIVE
Notifies takeover
Confirms takeover
Informs about current state
Transfers control to human
Core action
Feedback action
Order between actions
[...] Action condition



Opportunity - Prerequisites:

human.location = InDriverSeat; human.hands = OnTheWheel; radio.volume = Low

Willingness - Predisposition:

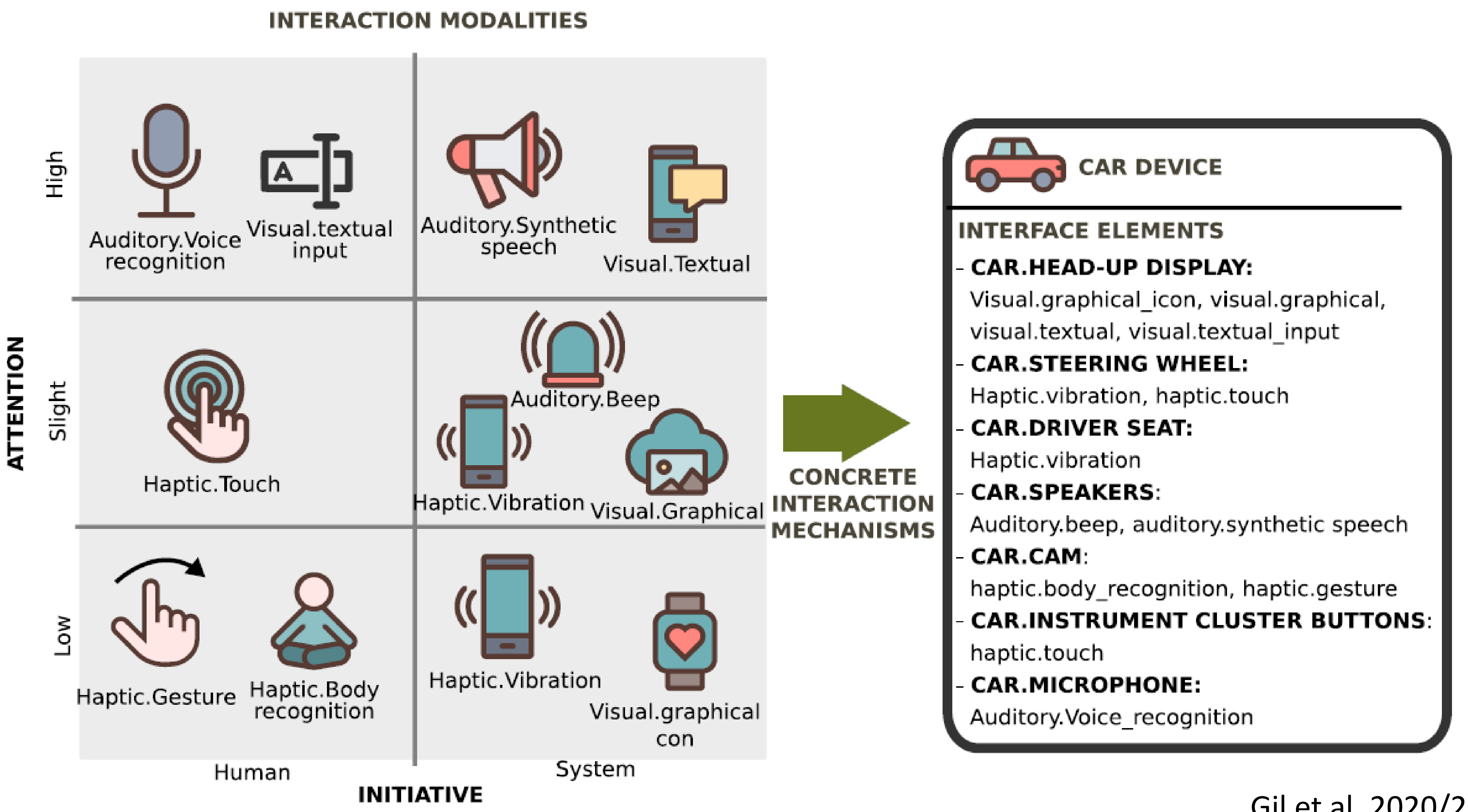
human.attention = High; human.stress = Low;

Capacity – Skill/Ability/State

human.emotion = Positive; human.training = High

Gil et al. 2020

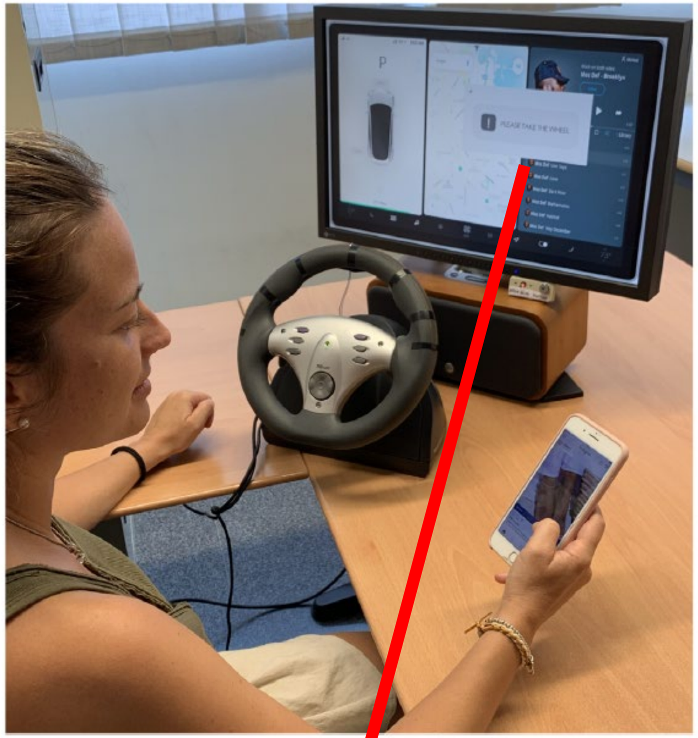
Obtrusiveness Level for HiL Action

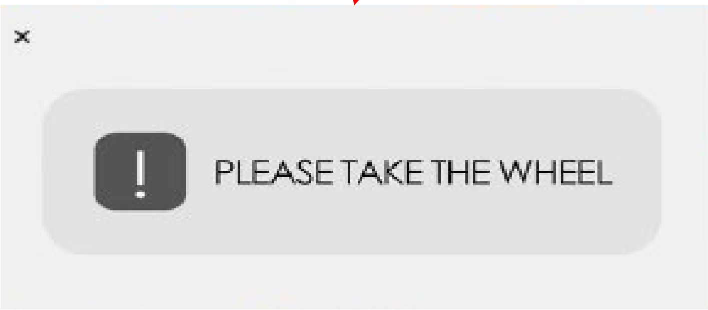


Gil et al. 2020/2022

Evaluation of HiL-CPS

- **Iterative prototyping**
- **User-based evaluation of prototypes**
  - Different fidelity: Wizard of Oz;
  - Feedback in different modalities: combination of text, audio, and visual
  - A range of scenarios: attentive/inattentive driver; system/human initiative
- **Dimensions**
  - Usability: Effectiveness, Efficiency, Satisfaction
  - User Experience: Excitement, Frustration, Pleasure
  - **Trust**
- **Instruments**
  - Quantitative: Questionnaires; Psycho-physiological data; Logging; Time; Errors
  - Qualitative: Interviews; Video-recording;





Implications for Trust …

References

- Fisher, M., Dennis, L., & Webster, M. (2013). Verifying autonomous systems. Communications of the ACM, 56(9), 84-93
- Hoffman, R.R., Mueller, S.T., Klein, G., and Litman, J. (2018). "Measuring Trust in the XAI Context." Technical Report, DARPA Explainable AI Program.
- Hoffman, R. R., Mueller, S. T., Klein, G., & Litman, J. (2018). Metrics for explainable AI: Challenges and prospects. arXiv preprint arXiv:1812.04608.
- Lage, I., Ross, A., Gershman, S. J., Kim, B., & Doshi-Velez, F. (2018). Human-in-the-loop interpretability prior. Advances in neural information processing systems, 31.
- .Lam, C. P., & Sastry, S. S. (2014, December). A POMDP framework for human-in-the-loop system. In 53rd IEEE conference on decision and control (pp. 6031-6036). IEEE.
- Nunes, D. S., Zhang, P., & Silva, J. S. (2015). A survey on human-in-the-loop applications towards an internet of all. IEEE Communications Surveys & Tutorials, 17(2), 944-965.
- Wu, X., Xiao, L., Sun, Y., Zhang, J., Ma, T., & He, L. (2022). A survey of human-in-the-loop for machine learning. Future Generation Computer Systems.