



DeepTech Track @ DevFest

Chennai 2026



October 17, 2026 / GDG Chennai DevFest

When AI Leaves the Lab : Building Physical AI People Can Trust

Pulkit Verma

IIT Madras

pulkitv@cse.iitm.ac.in / pulkitverma.net



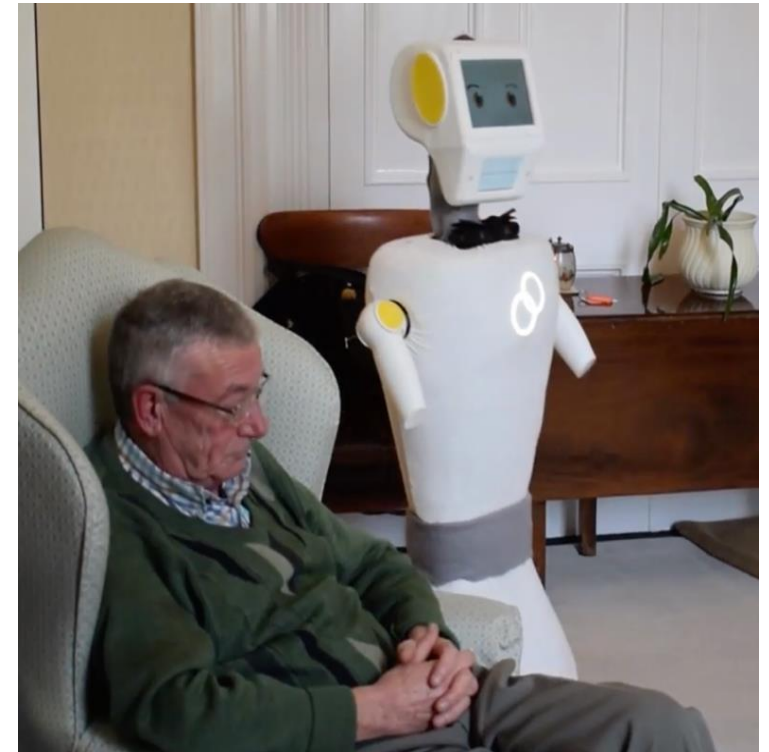
Robots that adapt to **your** home, **your** tasks, **your** preferences



it cleans

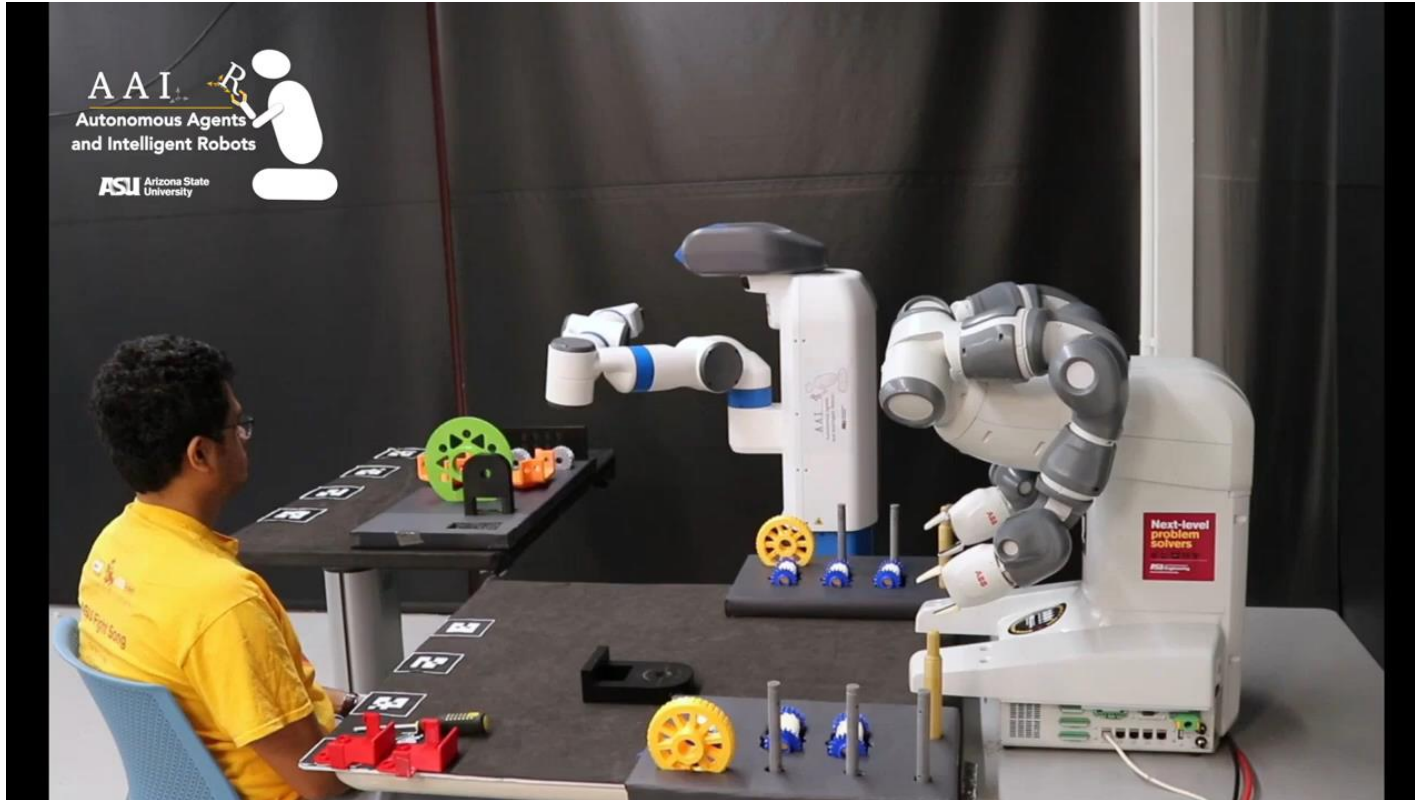


it cooks



it cares

You don't program it. You hand it a task.



Taskable AI systems

take a sequence of actions over time to get something done for a user.

They learn, they plan, and they keep changing after they ship.

The user is not the designer.

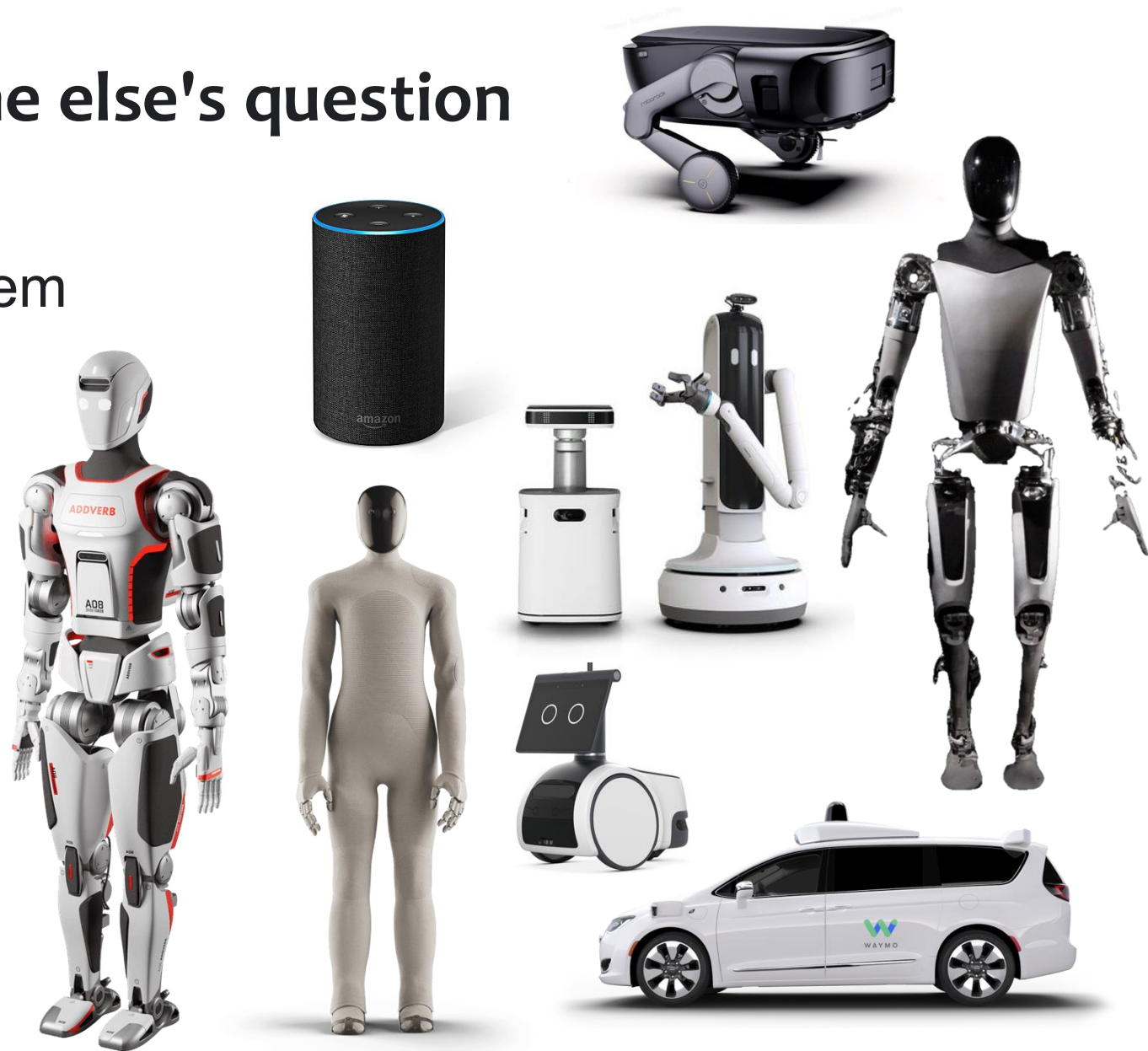
A user gives a task. The robots work out how.

Benchmarks answer someone else's question

A benchmark tells you how a system did on **someone else's** task.

It rarely tells you what it can **safely do for you.**

You will hand these systems many tasks, in many places, for years. Someone who is not the designer has to decide what to trust them with.



the question for this talk

When is it safe
for **THIS** user
to operate **THIS** robot
on **THIS** task
in **THIS** environment ?

Not "is the robot safe". Safe for whom, doing what, where.



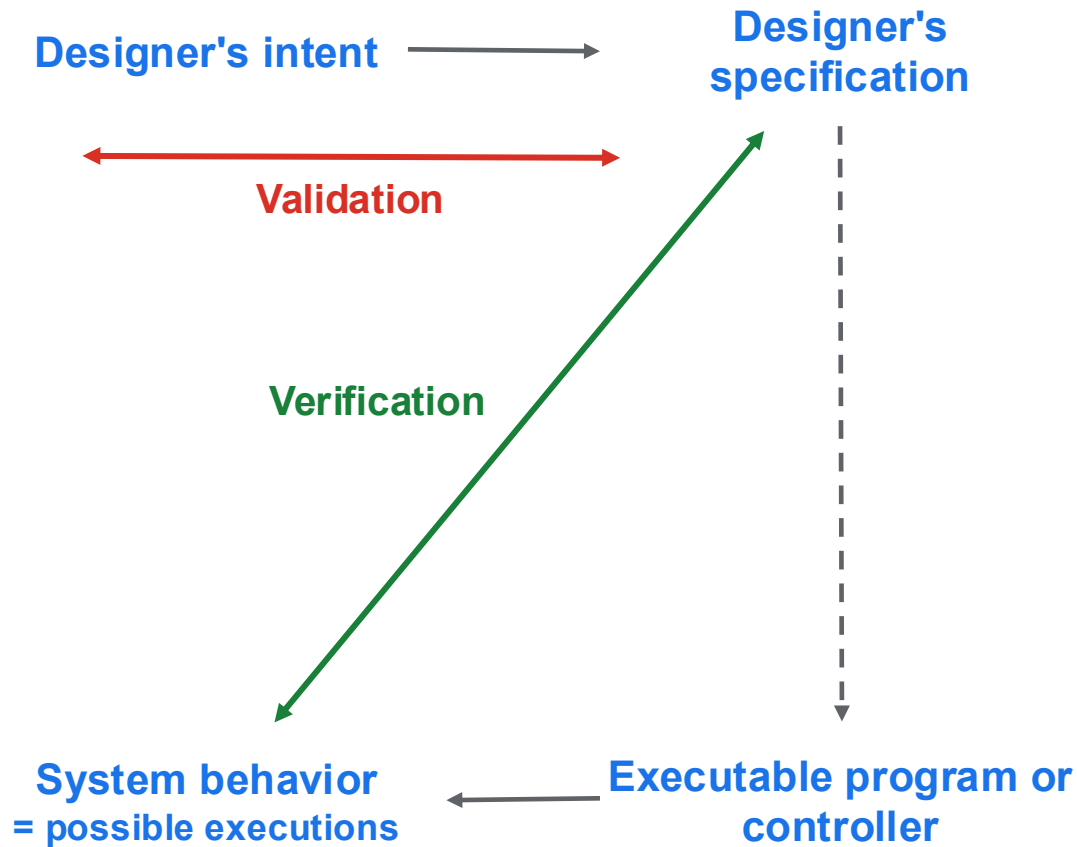
01

The designer leaves the room

Why the way we have always checked software does not fit a robot in your home



How we have always checked software



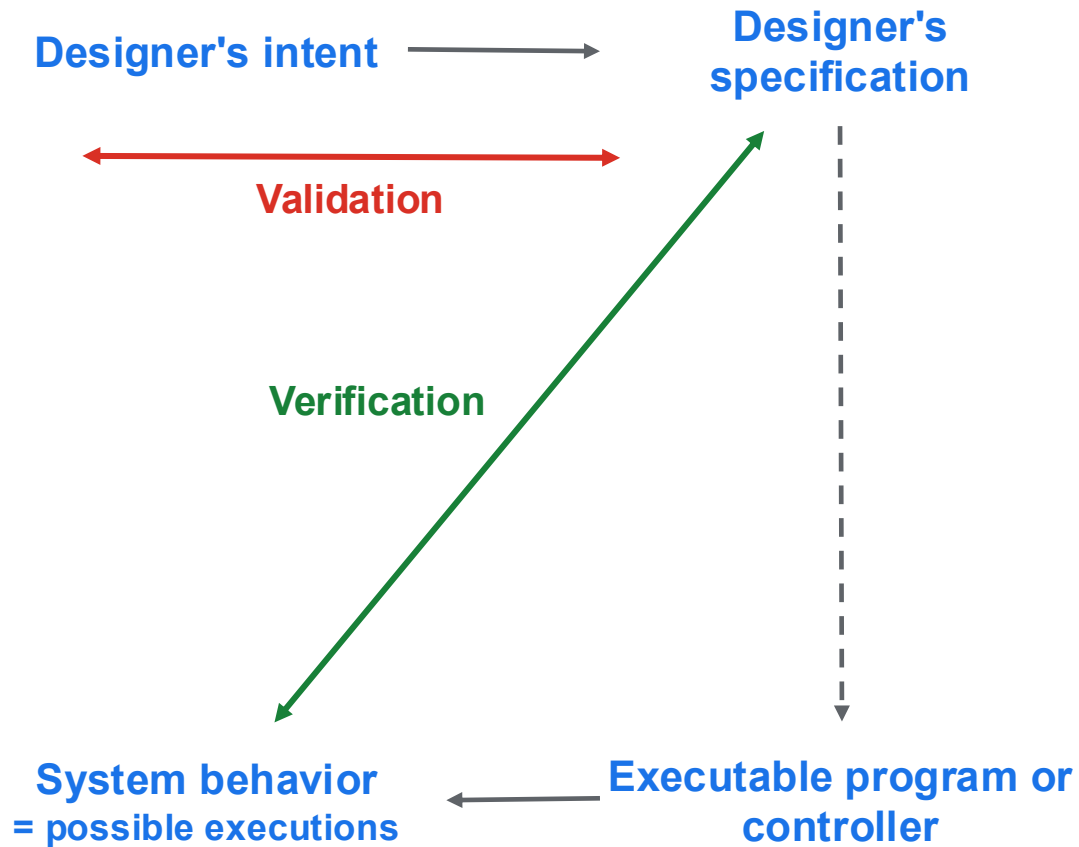
Verification asks if we built the thing right.

Validation asks if we built the right thing.

Input is the program, a model of the environment and a safety property.

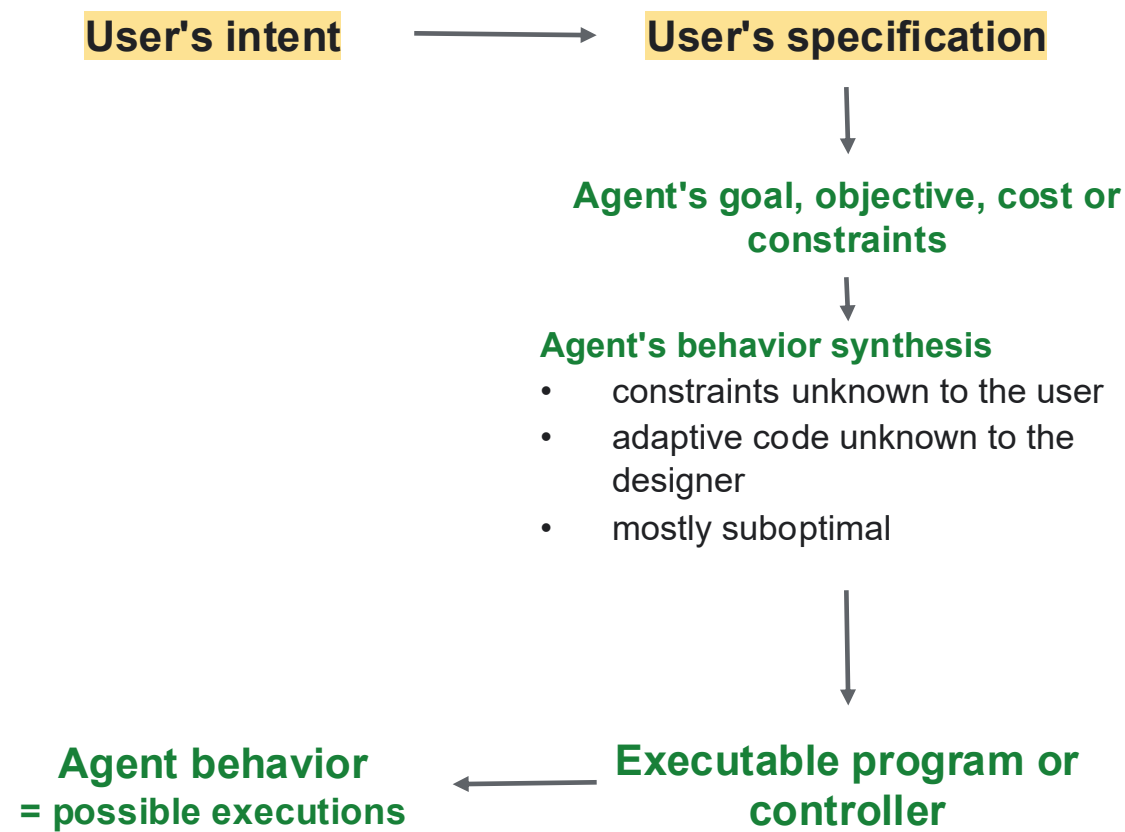
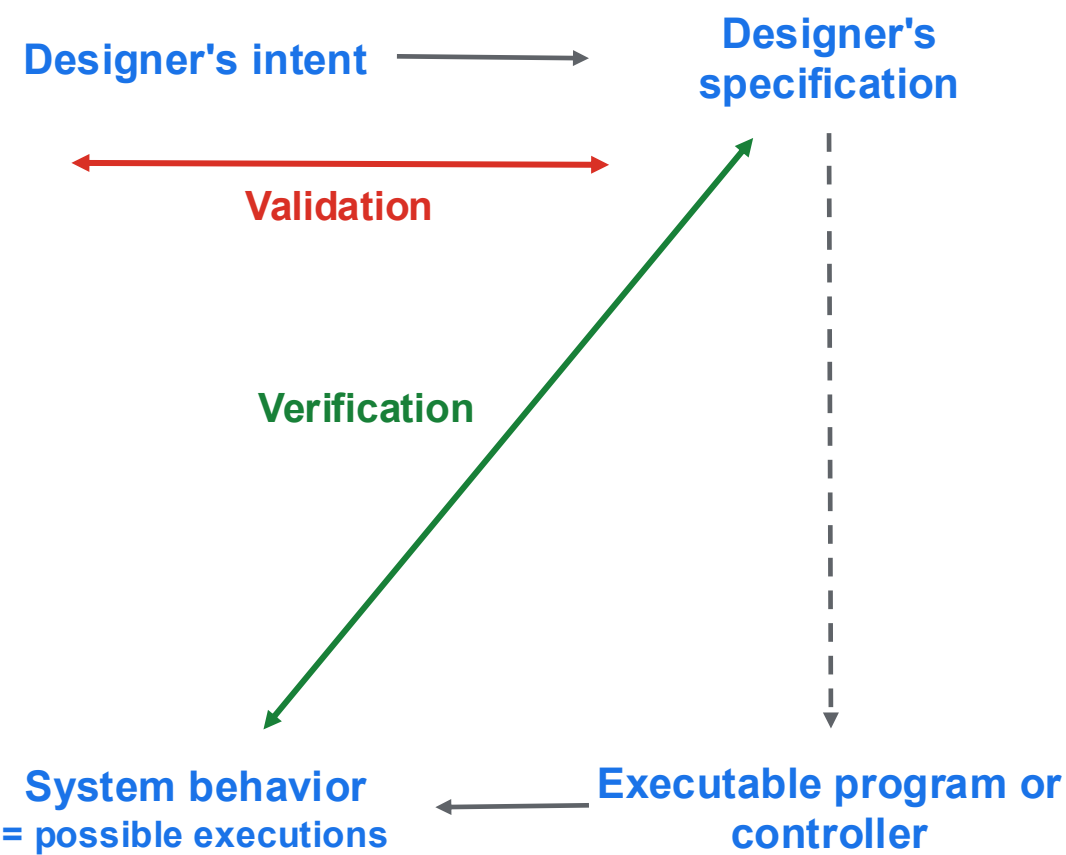
The designer sits at the center of both.

Now the user brings the task



Written after deployment, by someone the designer never met.

And the robot works out the rest



The designer is out of the loop



The designer is out of the loop.

The behavior you actually experience is synthesized at run time, in your home, for your task.

How would a user know it is safe to hand over this task?



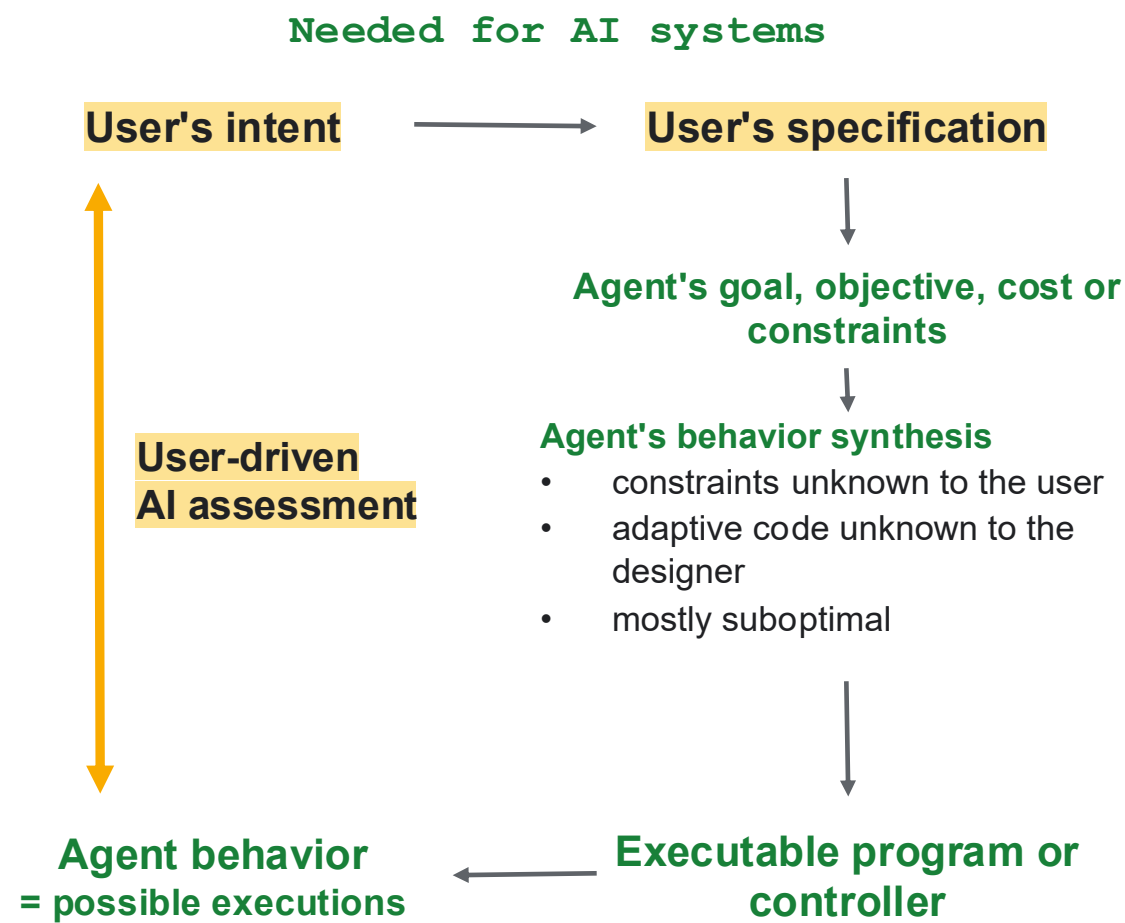
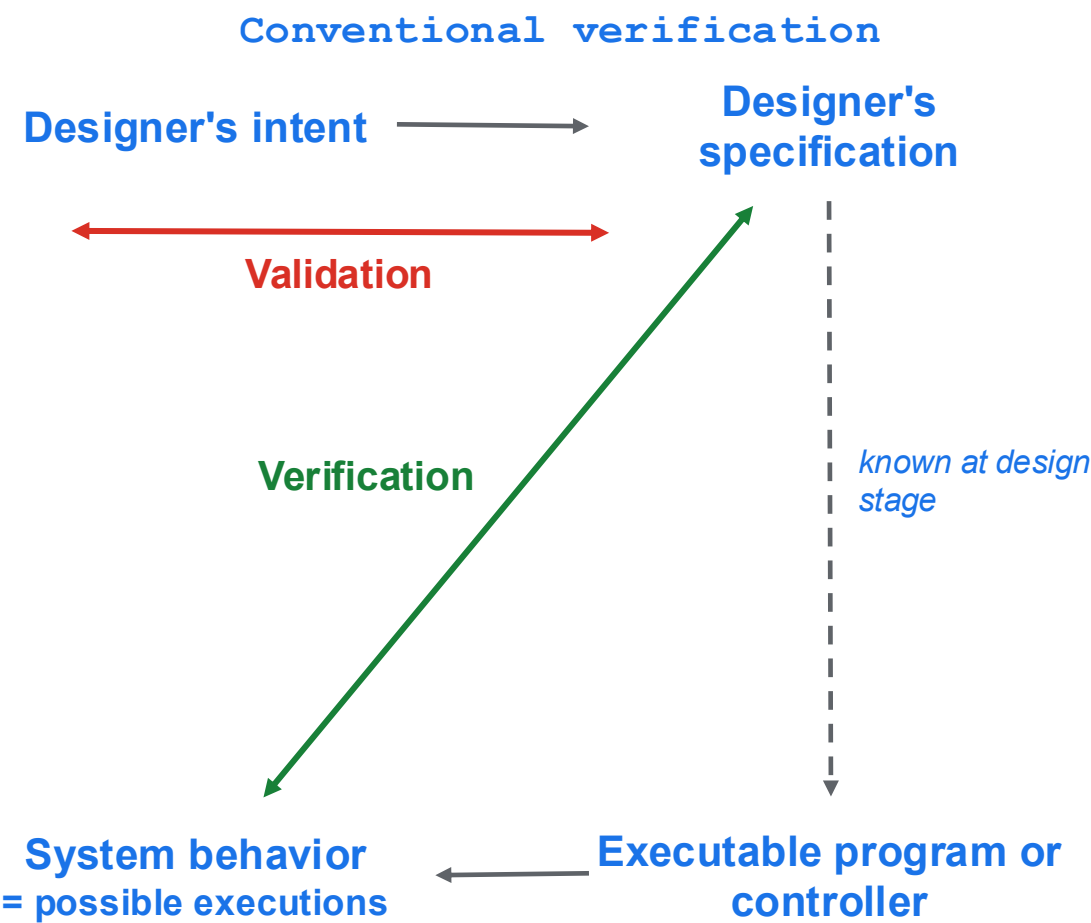
↓
Agent's goal, objective, cost or constraints

↓
Agent's behavior synthesis

- constraints unknown to the user
- adaptive code unknown to the designer
- mostly suboptimal



Assessment has to start from the user



02

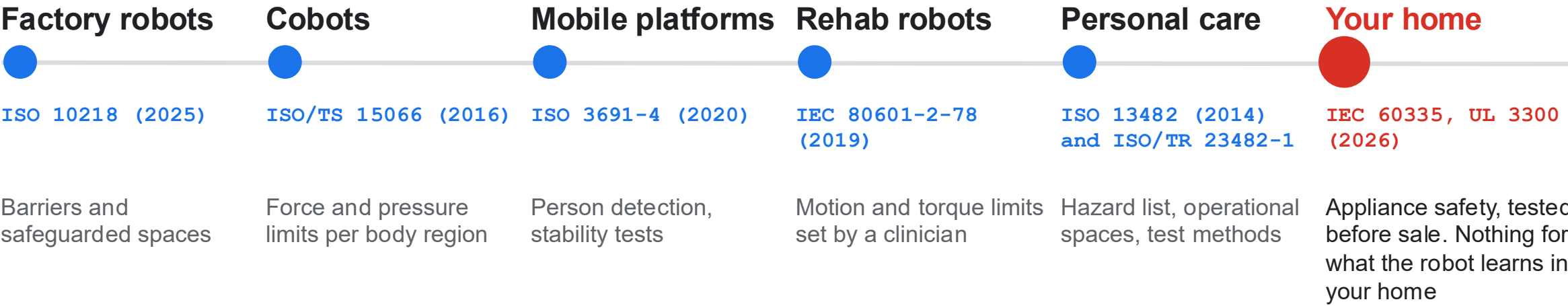
Who models safety at home?

What robot standards and AI standards assume, and who they forget



Safety standards walked the robot out of the cage

??



Every one of them fixes safety at **design time**, and assumes someone already knows the task, the space and the operator.

Who draws these lines in your living room?

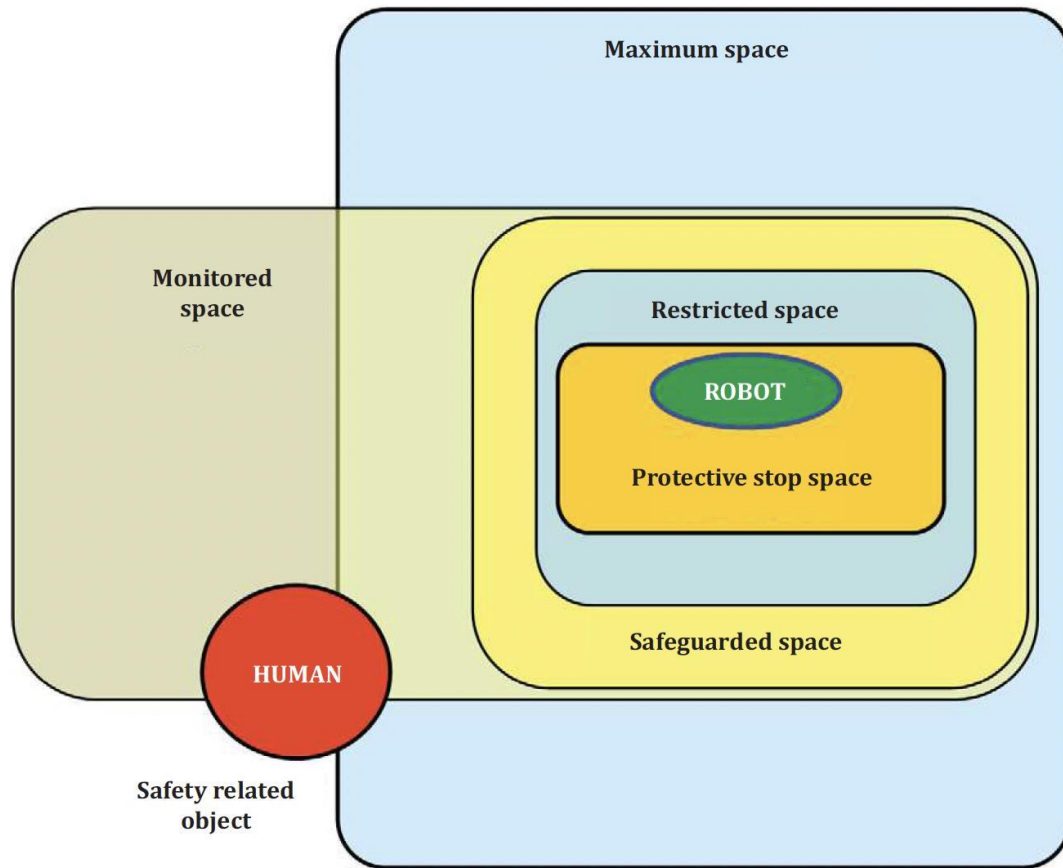


Figure B.2 — Operational spaces of a personal care robot with manipulator

ISO 13482 gives a personal care robot five kinds of space

maximum
restricted
monitored
safeguarded
protective stop

In a factory, a safety engineer draws them.
In a clinic, a therapist sets the limits.

In your home, nobody does.

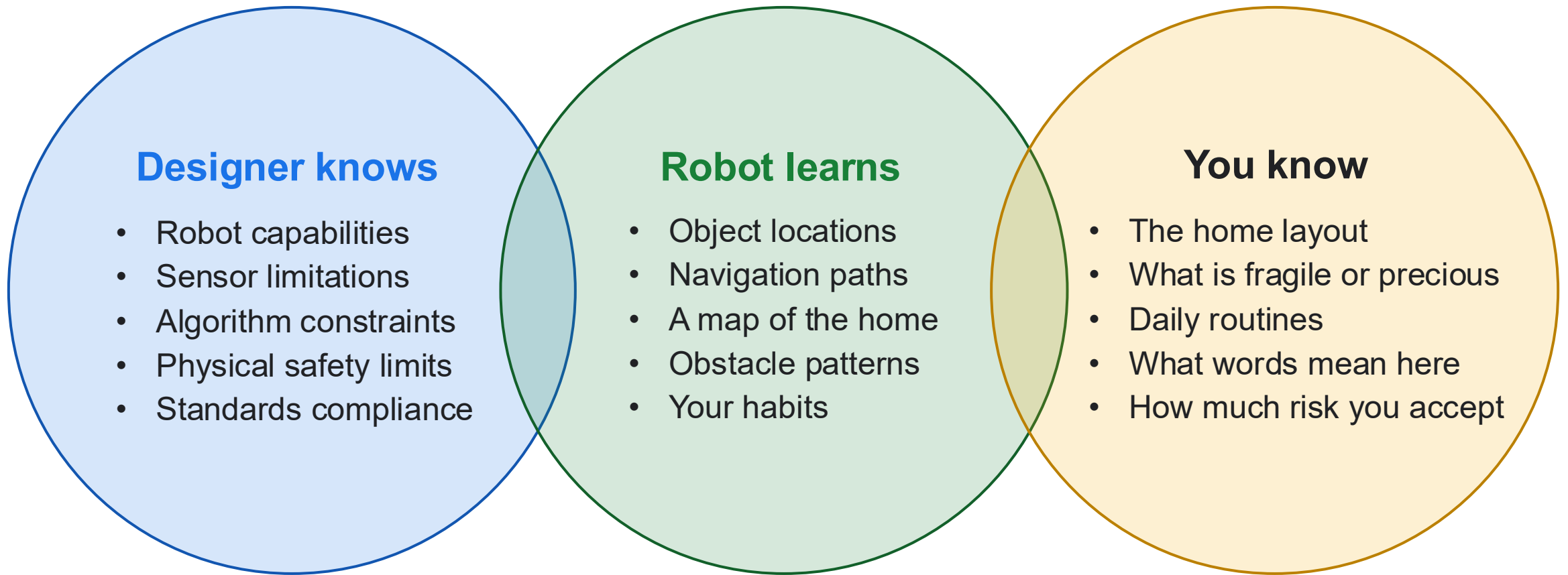
Source ISO 13482 (2014), Annex B

AI standards manage risk for a company, not your home

standard	what it gives you	what it misses at home
ISO/IEC 22989	Shared terms and stakeholder roles	Gives terms, not an assessment method
ISO/IEC 23894	AI risk management process	The organization owns the risk
ISO/IEC 42001	AI management system	Certifies the company, not the robot in one home
ISO/IEC 42005	AI system impact assessment	Done by the organization, without the end user
ISO/IEC TR 5469	AI inside safety functions	Assumes the safety function is defined in advance
ISO/IEC TS 8200	Controllability and handover	Does not say how a user expresses intent
ISO/IEC TS 6254	Explainability objectives	No format tied to the user's own concepts

At home, the customer, the user and the person who can get hurt are often **the same person**.
The EU AI Act leaves that private user largely outside its scope.

Safe operation needs knowledge that nobody holds alone



A complete operational model needs all three.

“Tidy the living room carefully.”

	You mean	The robot reads	The designer verified
carefully	Don't disturb my papers on the floor	Move at a lower speed	Speed limits and protective stop distances
living room	The room with the couch	The room labelled “living” in its map	Navigation in generic indoor spaces
tidy	The antique rug is precious	Papers and rug are unknown objects to avoid or ignore	Generic hazard scenarios, nothing about this home

The robot can satisfy every design constraint **and still drag the sofa across the rug.**

Safety lives in the union of three models

$$M_{designer} \cup M_{user} \cup M_{robot}^t = M_{complete}$$

designer

Available only at design time. Standard tests verify it, but the maker never sees your home.

robot

Always incomplete, always changing. What held at time t may not hold after the next update.

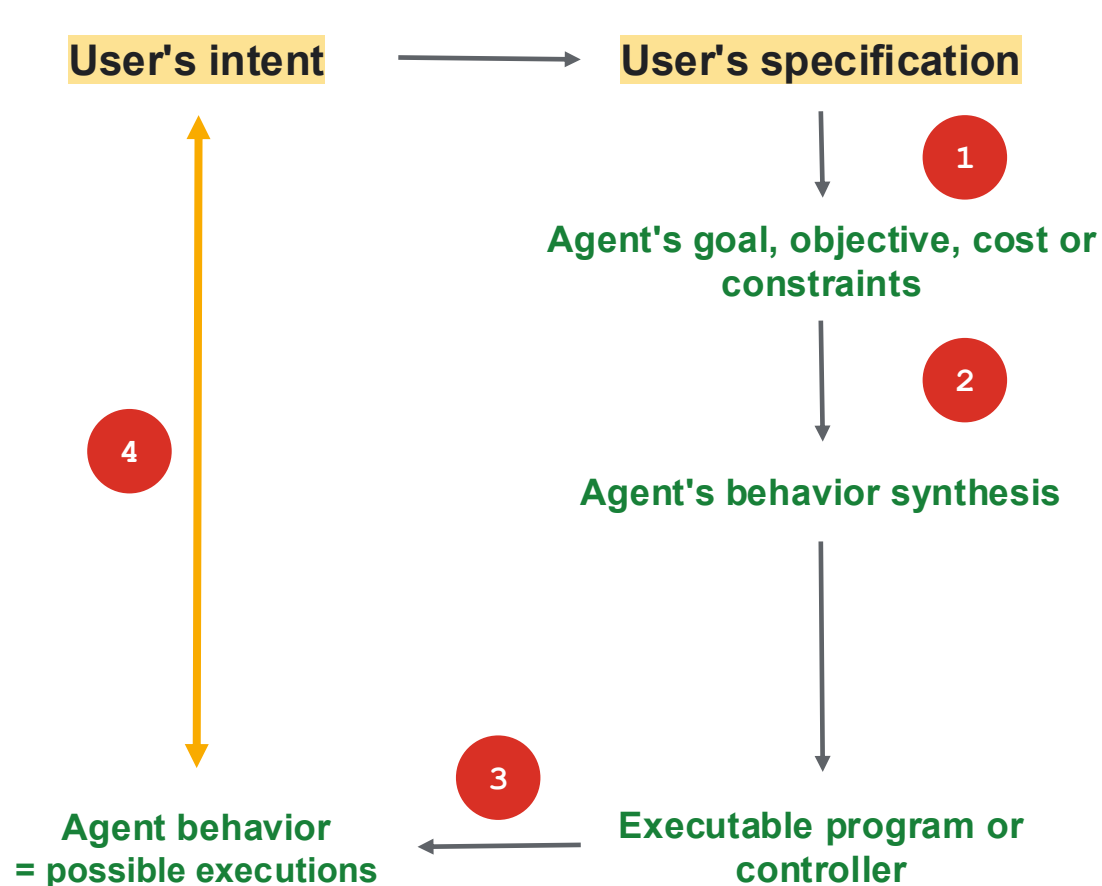
user

Never written down. “The papers stay put” has no counterpart in a model of speeds and forces.

complete

No party holds it at any time. Assessment has to be a process that combines all three.

Four places where your task falls through



Some safety rules only exist in one home

Standards ship a fixed list

85

hazard scenarios in ISO 13482

48

safety requirements in ISO 10218

17

test procedures in ISO/TR 23482-1

Your home needs rules like these

Do not enter the nursery between 13.00 and 15.00

Do not drag anything across the rug

The papers on the floor stay where they are

They depend on the task and the context, so they have to be **generated** from both and **checked while the robot runs**. No standard in either family does this yet.

03

Ask the black box

Discovering, in your own words, what an agent can and cannot do



Ask the black box questions, in your words

```
at(p0,cell_6_3)
clear(cell_0_0)
door_at(cell_9_2)
next_to(m0)
alive(m0)
key_at(9_4)
```



Concepts you understand



A description of its capabilities, in your words



Personalized AI Assessment Module (AAM)

query



response



Black-box AI



simulator

any internal implementation doesn't know your vocabulary

Get back what it can do, in your words

```
at(p0,cell_6_3)
clear(cell_0_0)
door_at(cell_9_2)
next_to(m0)
alive(m0)
key_at(9_4)
```

Now I understand
what it can do!

Concepts you
understand



A description of
its capabilities,
in your words



Personalized AI
Assessment Module (AAM)

query

response



Black-box AI



simulator

any internal
implementation
doesn't know your
vocabulary

A capability you can read out loud

```
(:action open-door
  :parameters (?l1)
  :precondition (and (has_key)
                     (player_at ?l1)
                     (door_adjacent ?l1))
  :effect (probabilistic
    0.95 (and (door_open))
    0.05 (and (not (has_key))
              (game-over))))
```

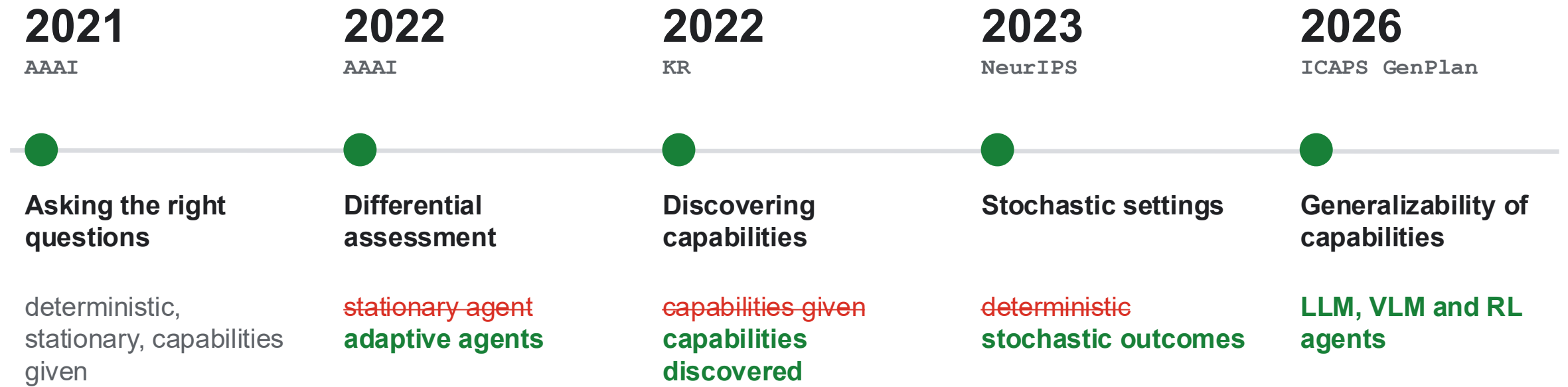
You can open the door at a location if **you have the key, you are standing there, and the door is next to it.**

Then **95 times out of 100 the door opens.**

5 times out of 100 you lose the key and the game is over.

Learned as PDDL, a planning language. Conditions, effects and odds, all in the user's vocabulary.

Each step dropped an assumption



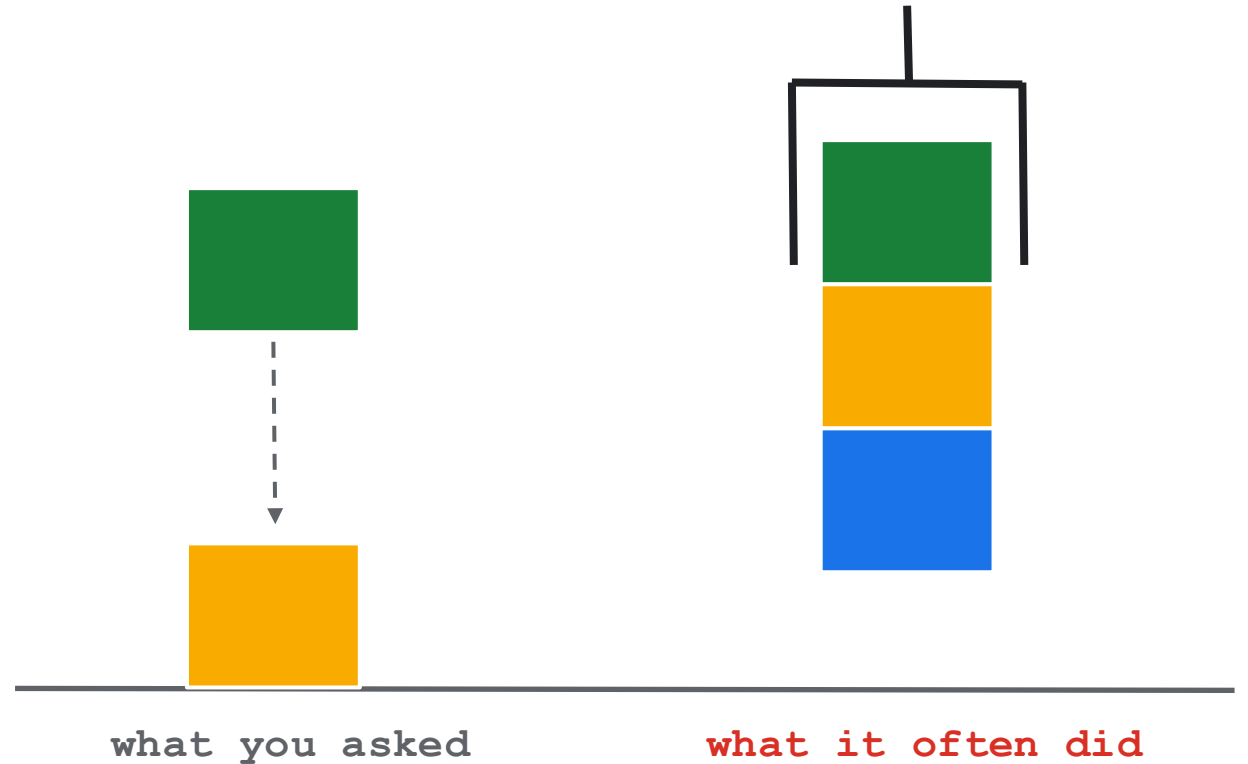
Verma, Marpally, Srivastava (AAAI 2021, KR 2022).
Nayyar, Verma, Srivastava (AAAI 2022).
Verma, Karia, Srivastava (NeurIPS 2023).
Bramblett, Karia, Ciotinga, Verma, Choi, Srivastava (2026).

Asked to stack one block, it lifted the whole tower

6%

of the time, a simulated SayCan-style mobile manipulator actually put the green block on the yellow one.

That side effect is exactly what you need to know before you hand over a task.



Bramblett et al. 2025, assessing LLM-based, VLM-based and RL agents

An explanation can sound right and still be wrong

A bird classifier credits “**red plumage**” for every cardinal. Its decision is really driven by “**has a crest**”.

A driving policy lights up “**passenger pickup**” while that concept has no effect on what the car does.

If a robot explains itself to you in your words, the words have to be the reason.

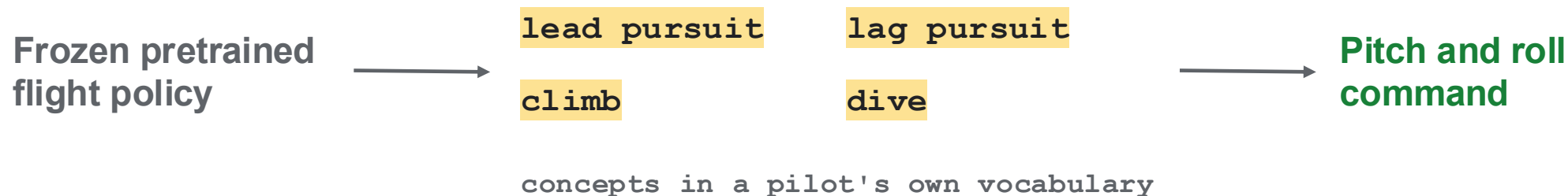
97.6%

of bird concepts predicted correctly by a standard concept bottleneck model

about 6%

of its decision actually flows through those concepts, close to random chance

Train the explanation to be the cause



Necessary. Swap out a concept that is not active, and the action should not change.

Sufficient. The active concepts alone should rebuild the action.

Independent. Each concept's effect adds up cleanly, with no hidden mixing.

92 to 96%

of test cases route over 95% of the decision through the right concepts. The closest baseline manages 32 to 59%.

04

When it still goes wrong

Liability, lessons from aviation, and a hybrid path forward



Who is liable when the robot learned it at your place?

close analogy

weak analogy



Autonomous vehicles

Liability shifts toward the maker as autonomy rises.

But a home has many “drivers”, no licence, and you teach the robot.

Medical devices

A trained physician sits between the device and the patient.

At home, you are both the physician and the patient.

Smart home devices

Plain product liability. A failed thermostat costs comfort.

A failed robot can injure someone.

Learned behavior is not a manufacturing defect, not a design defect, not a warning defect.

Causation is spread across the maker, the home and the task.

Aviation separates two things we keep bundling

Investigation

standardized and neutral

- Same protocol worldwide
- Fact finding, not blame
- Root cause analysis
- Shared database of failure patterns
- Recommendations for change

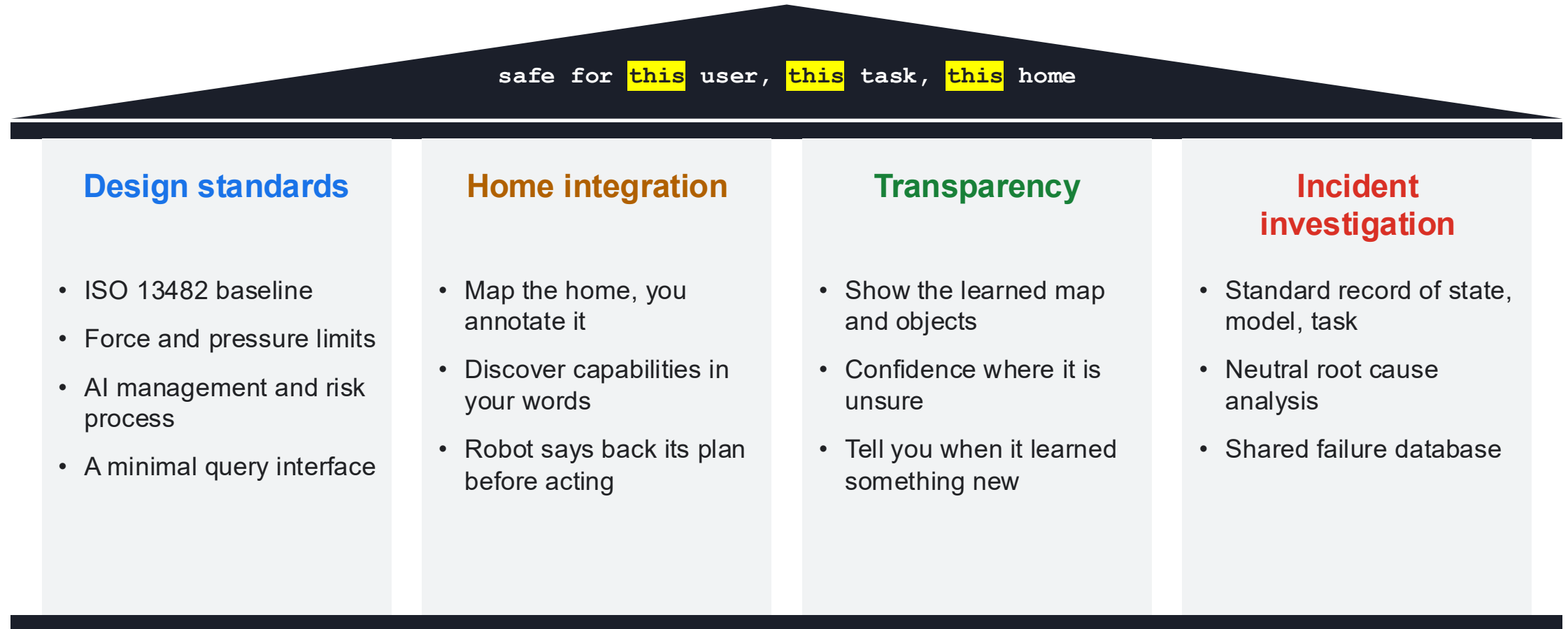
Responsibility

decided by context

- Varies by jurisdiction
- Insurers and courts
- Regulators
- Equipment and training changes

ICAO Annex 13. Search for blame too early and nobody shares the data that explains what happened.

Keep the standards. Build the missing floors on top.



Current standards are necessary. They are not sufficient.

What you can build into your system on Monday

01

Expose a query interface, not your weights.

Let a third party ask your agent “what happens if”.

02

Describe capabilities in your user's vocabulary.

Conditions, effects and odds, not a manual written before it learned anything.

03

Say the task back before acting.

“I will tidy the room with the couch. I will not move the papers.”

04

Explain in concepts that actually drive the decision.

Show what it learned and where it is unsure, and test that the explanation is causal.

05

Keep a record an investigator could learn from.

State, model, task and user annotations, without leaking private life.

Capable is not
enough.

Physical AI has to be
human compatible.

Thank you

pulkitv@cse.iitm.ac.in
pulkitverma.net

with



Siddharth
Srivastava



Rushang Karia



Shashank R
Marpally



Julie A Shah

and Daniel Bramblett, Rashmeet K Nayyar, Georgios Fainekos, Adrian Ciotinga, YooJung Choi, Josh Rountree, Oswin So, Chuchu Fan

supported by

