Human in the Loop
learning and control

Nikita Lukianets,
Founder @Open Ethics
https://openethics.ai/

"
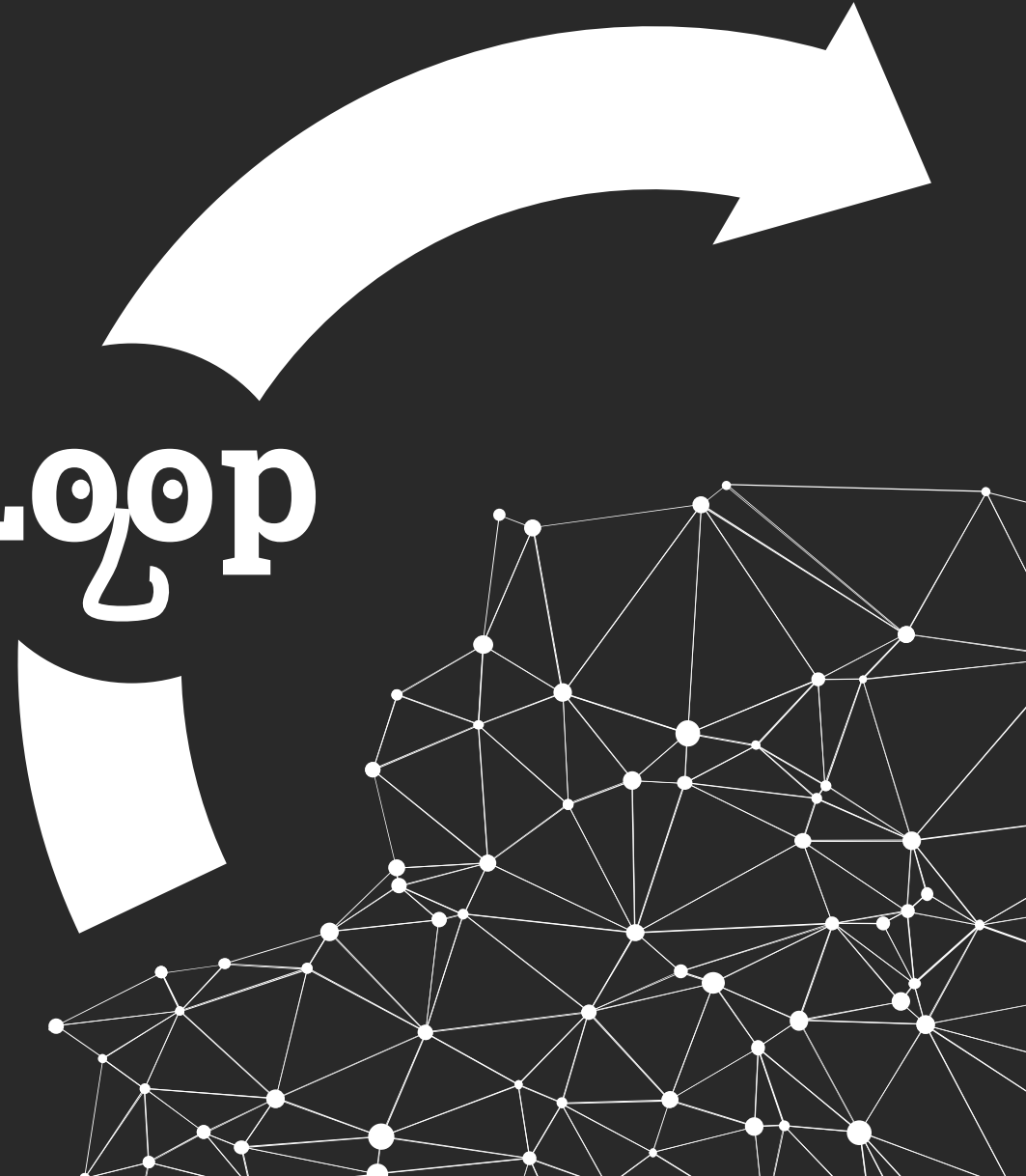
A system that makes it possible for a computer to learn from experience, adjust to new inputs, and perform tasks commonly associated with human intelligence.
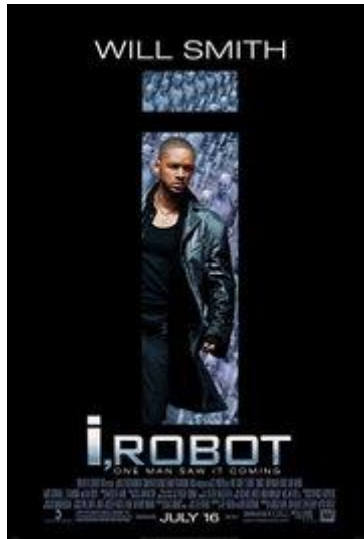
1. Perceive environment + **2. Act to succeed at goal**

Intelligence...

AI Ethics is NOT about this

this?

Or this?

What is the threat?

Building decision models

Perceive	Decide	Act	Evaluate
What information do we have/need to make decisions?	How do we decide based on received information?	What do we do once we have our decision?	How do we evaluate the quality of our decision?
(better)	(better)	(better)	(better)

Labeling: huge amounts of work...

SML training depends on data annotated by subject-matter experts (Lawyers, Linguists, Doctors)

Benefits and Risks

Highlight non-obvious relationships in big data

Augment human decision-making

Save time

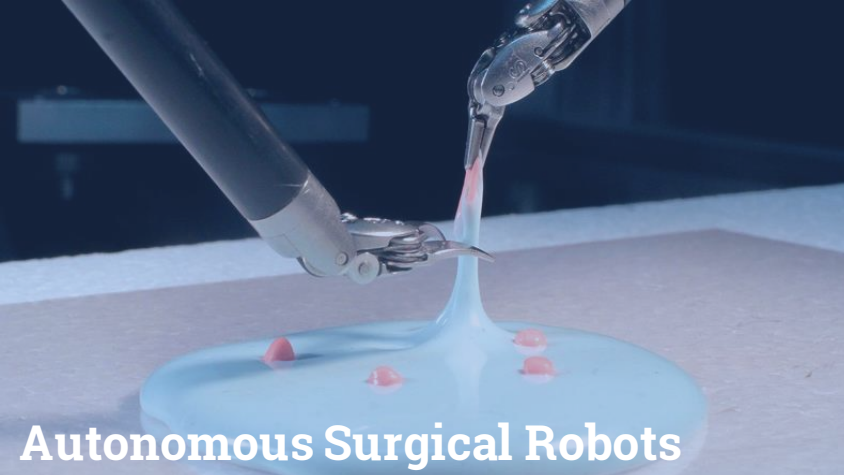Learn operational protocol or user **personal preferences**

Lower **stability against adversarial input**

High **false-positive** rates

Too much **reliance on humans** for operations

Lower **interpretability**

Carbon Black, Beyond the Hype: Artificial Intelligence, Machine Learning and Non-Malware Attacks Research Report 2017
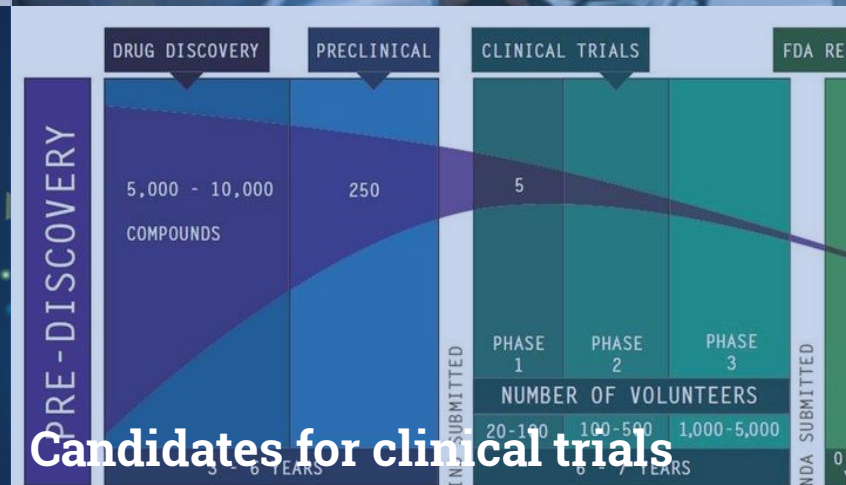
Autonomous Surgical Robots
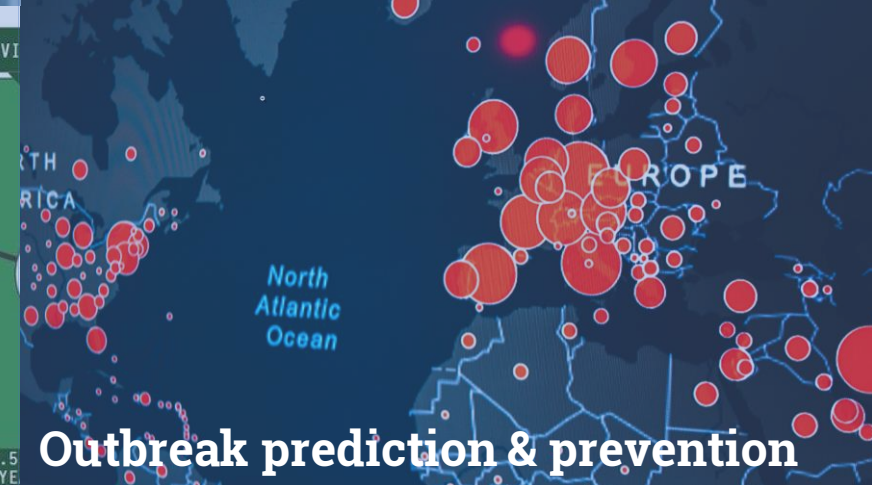
Automated diagnostic tools
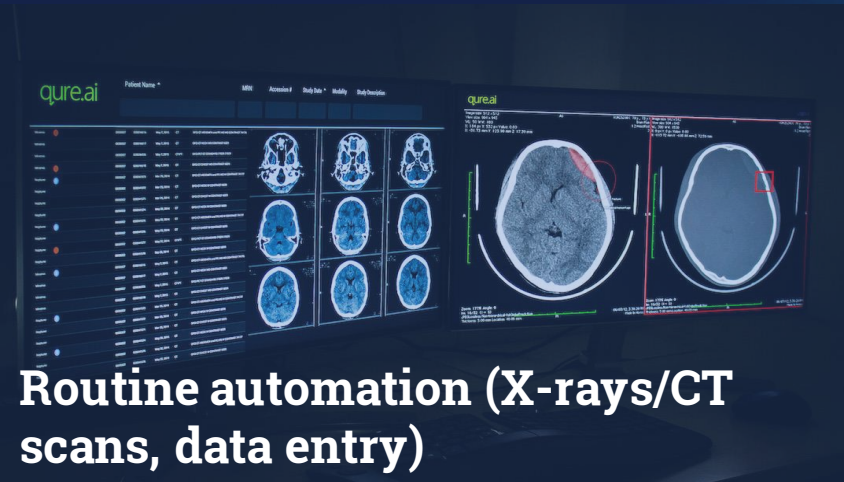
Personalized treatment
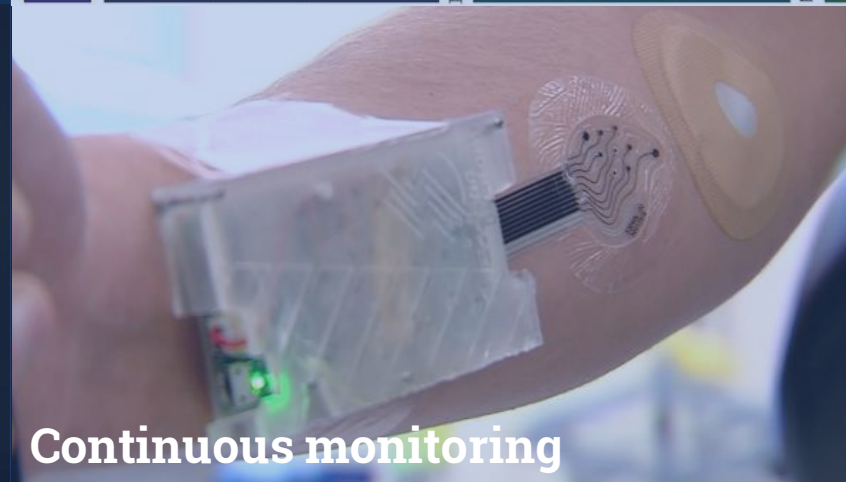
Drug discovery
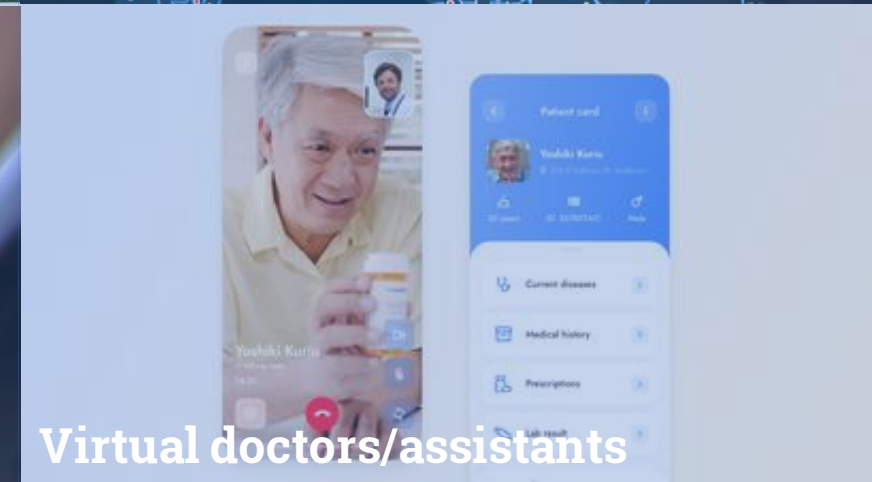
Candidates for clinical trials

Outbreak prediction & prevention
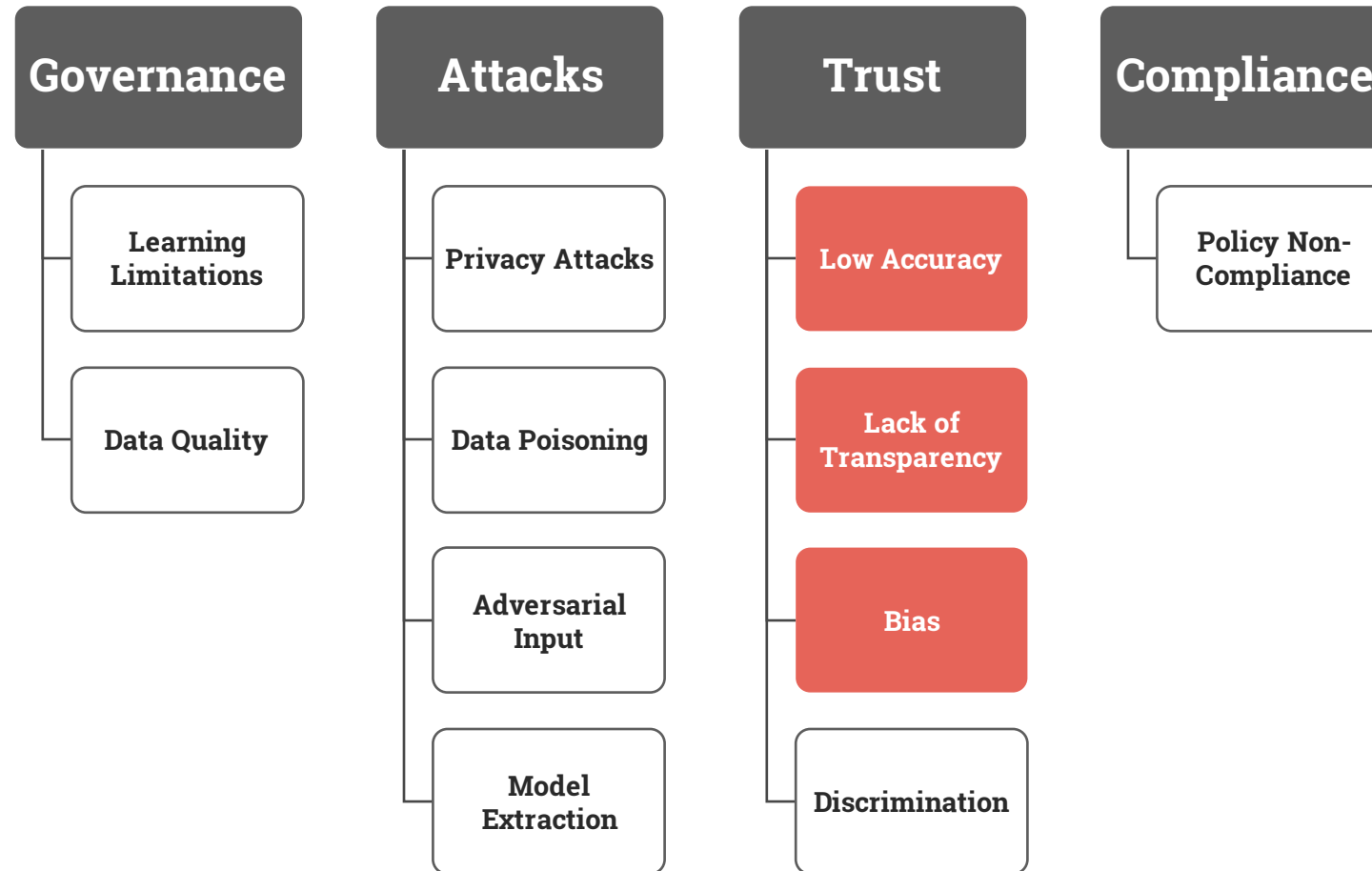
Routine automation (X-rays/CT scans, data entry)

Continuous monitoring

Virtual doctors/assistants

AIRS AI Risk Categorization

Governance
- Learning Limitations
- Data Quality

Attacks
- Privacy Attacks
- Data Poisoning
- Adversarial Input
- Model Extraction

Trust
- Low Accuracy
- Lack of Transparency
- Bias
- Discrimination

Compliance
- Policy Non-Compliance

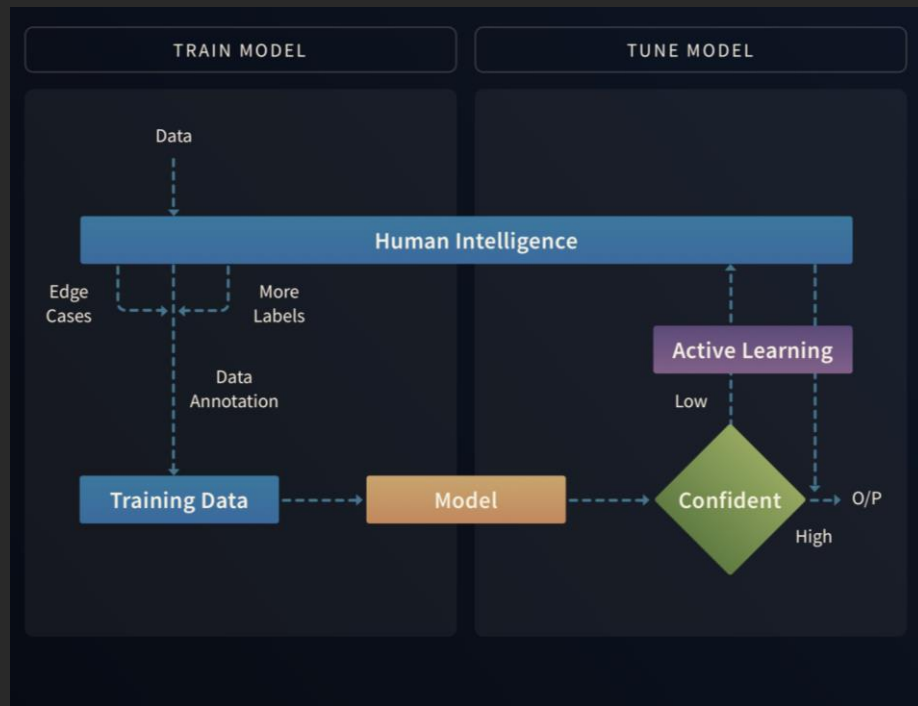Based on Wharton AIRS group report on AI Risk and Governance

Why to use Human-in-the-Loop ML?

- Making Machine Learning models **more accurate**
- Getting Machine Learning to the desired **accuracy faster**
- Making **humans more accurate**
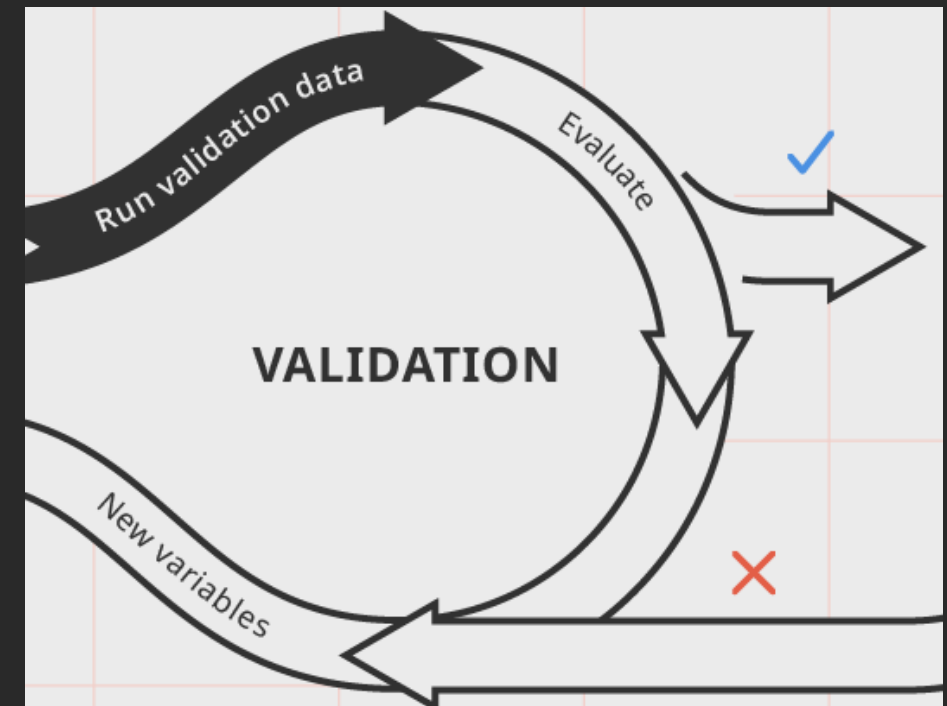- Making humans **more efficient**

Learning

Focusing on smart annotation strategies

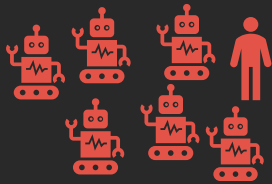**HITL = Supervised ML + Active Learning**

Control

Focusing on overall user experience and safety

Exploring validation configurations
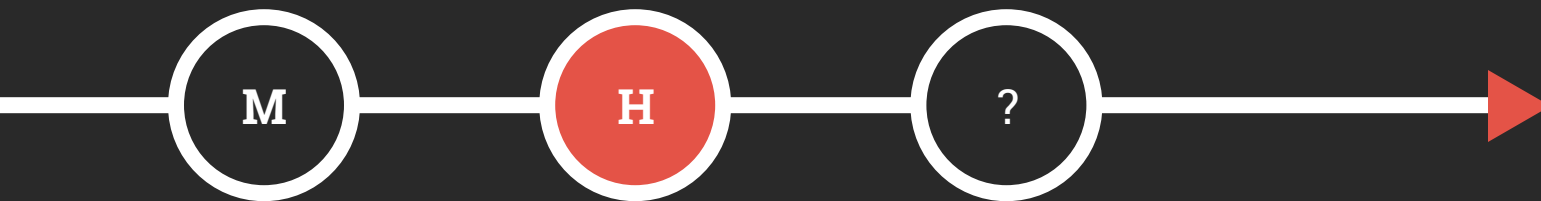
HITL control design

Swarm

Council

Crowdsourcing

Expert

M — H — ?

1. Humans confirm
2. Humans advise
3. Machines confirm
4. Machines advise

1. Mitigating Human Annotation Errors

Human ability to effectively label data depends on the learning sequence.

Quality can be improved by local changes in the **instance ordering** provided to the annotators.

Modeling and Mitigating Human Annotation Errors to Design Efficient Stream Processing Systems with Human-in-the-loop Machine Learning

Rahul Pandey[a,*], Hemant Purohit[a], Carlos Castillo[b], Valerie L. Shalin[c]

2. Going beyond validation of output

Human's role is elevated from simply evaluating model predictions to interpreting and even updating the model logic.

Providing **explainable or "middleware" solutions to validators** increase quality.

HEIDL: Learning Linguistic Expressions with Deep Learning and Human-in-the-Loop

Yiwei Yang[1], Eser Kandogan[2], Yunyao Li[2], Walter S. Lasecki[1], Prithviraj Sen[2]
[1]Computer Science and Engineering, University of Michigan - Ann Arbor
[2]IBM Research - Almaden, San Jose, CA
{yanyiwei, wlasecki}@umich.edu, {eser, yunyaoli, senp}@us.ibm.com

3. Cyber Physical Systems and HITL

HITL concept exhibits limitations due to the different natures of the systems involved.

It's proposed that besides human feedback loop, the **Bio-CPS validation** is needed.

Towards a theory for Bio–Cyber Physical Systems Modelling
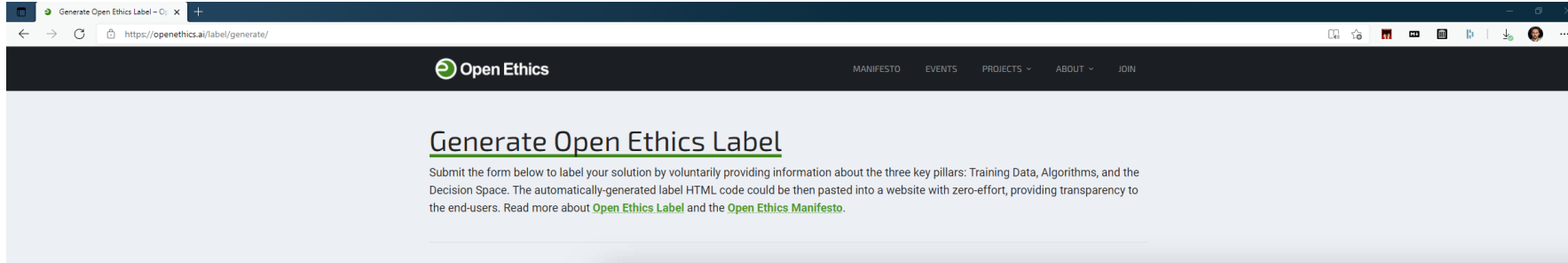
Didier Fass[1] and Franck Gechter[2]

4. Effective redundancy for Safety

Incorporate **human redundancy structures**, active and standby human redundancy, duplication and overlap of functions, and cognitive diversity.

Human redundancy in complex, hazardous systems: A theoretical framework

David M. Clarke

HITL in the self-disclosure process

Open Ethics Transparency Protocol

```json
1    {
2        "schema": {
3            "name": "Open Ethics Transparency Protocol",
4            "version": "0.9.2 RFC"
5        },
6        "data": {
7            "availability": "public",
8            "batch": 14032507,
9 >          "sources": [ …
11           ]
12
13       },
14 >     "algorithms": [ …
53       ],
54 >     "decision-space": { …
59       },
60       "validation": {
61           "critical": true,
62           "methods": [
63               {
64                   "name": "HITL",
65                   "objective": "To confirm the diagnosis by
                     human doctors before presenting to the
                     patient",
66                   "type": "council"
67               }
```
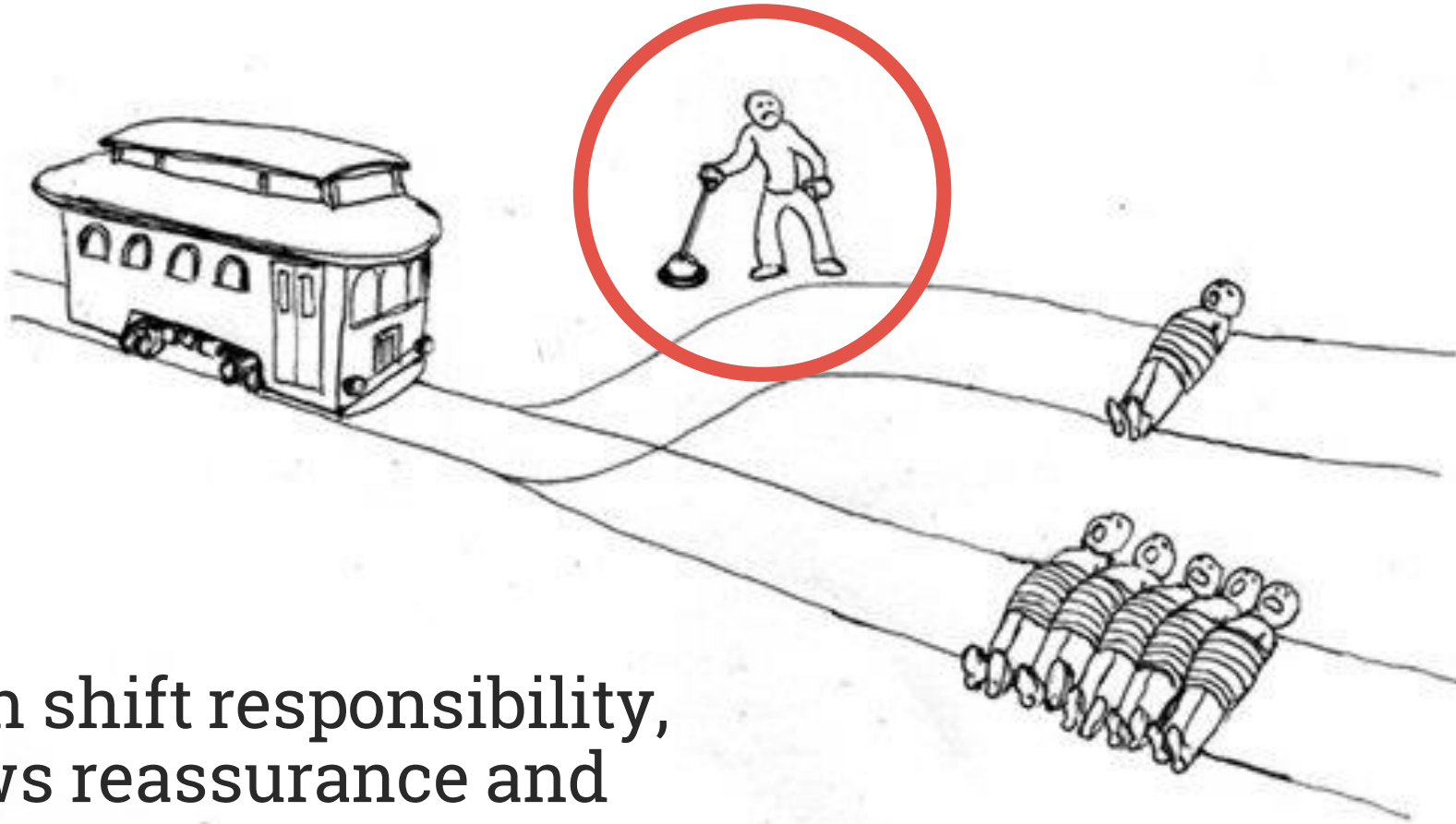
Every AI-powered product will have its passport to display "ethical" posture in both human and machine-readable ways.
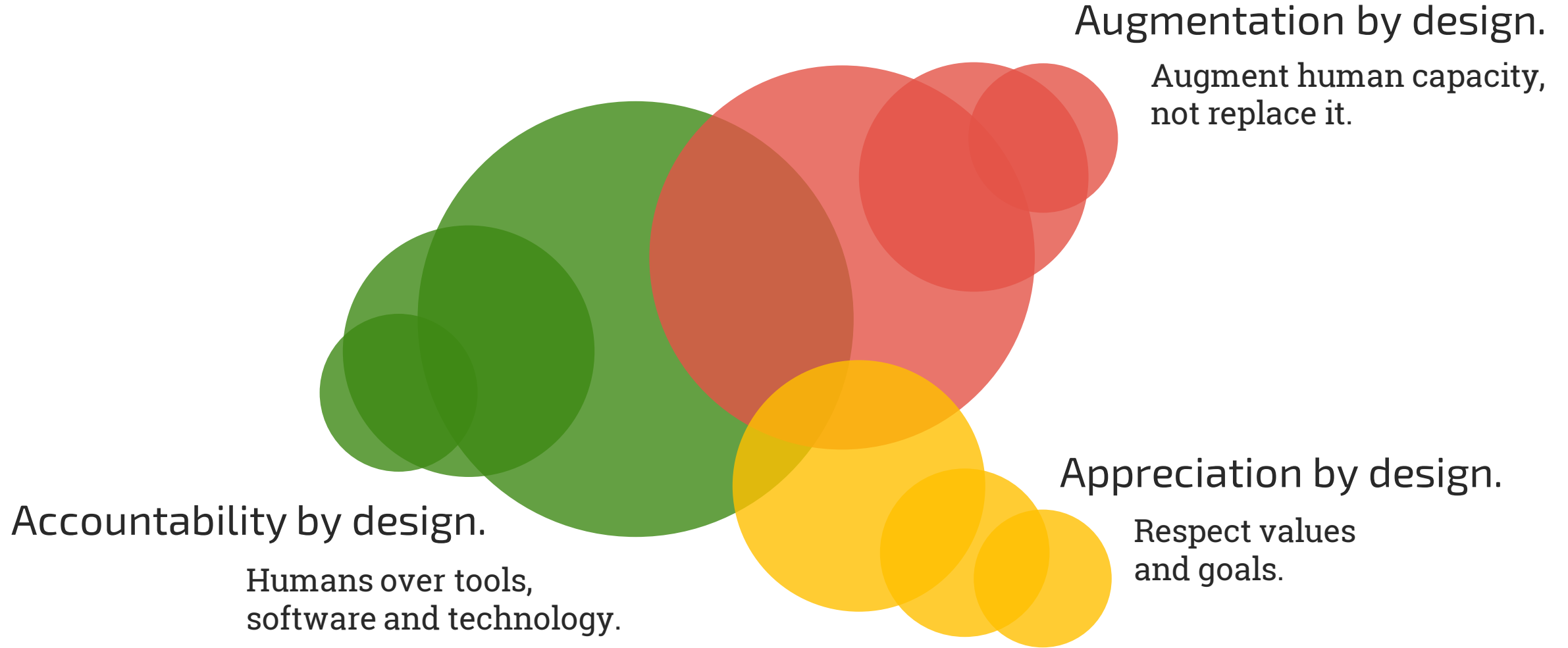
Does HITL setup requirement remove ethical concerns?

No, it can shift responsibility, but allows reassurance and two-way communication.

3A of AI ethics

Augmentation by design.

Augment human capacity, not replace it.

Appreciation by design.

Respect values and goals.

Accountability by design.

Humans over tools, software and technology.

Nikita Lukianets

n.lukianets@openethics.ai