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.

https://openethics.ai/taxonomy/
1. Perceive environment
2. Act to succeed at goal
AI Ethics is NOT about this

Nikita Lukianets @nikiluk

Quantitative Ethics - governance and ethics of AI decisions
What is the threat?

Or this?

this?
# Building decision models

<table>
<thead>
<tr>
<th>Perceive</th>
<th>Decide</th>
<th>Act</th>
<th>Evaluate</th>
</tr>
</thead>
<tbody>
<tr>
<td>What information do we have/need to make decisions?</td>
<td>How do we decide based on received information?</td>
<td>What do we do once we have our decision?</td>
<td>How do we evaluate the quality of our decision?</td>
</tr>
</tbody>
</table>

(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**

**Learn operational protocol or user personal preferences**

**Save time**

**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

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

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.
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.
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.
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, A. B.
HITL in the self-disclosure process
Open Ethics Transparency Protocol

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.

Accountability by design.
Humans over tools, software and technology.

Appreciation by design.
Respect values and goals.
Nikita Lukianets
n.lukianets@openethics.ai