Human in the Loop ML: Misunderstood — Summer Edition

Lazaros Toumanidis

Apr 04, 2026

Contents

> *The best way to understand a machine learning system is to imagine it without you in it. Then put yourself back.*

This is the companion to *Human in the Loop Machine Learning: Misunderstood* — same ideas, different pace.

One chapter at a time. A few questions to sit with. A puzzle, because puzzles are good for the same part of the brain that notices things AI gets wrong. No grades, no rubrics, no citations.

Grab a pencil. Dog-ears are encouraged.

How to use this:

Each chapter has three parts:

1. **Think about it** — a handful of questions. Not trick questions. Not exam questions. The kind you might wonder about on a walk.
2. **Spot the human** — a short scenario. Where is the human in the loop? What are they actually doing? Is it working?
3. **Word puzzle** — key terms hidden in a grid (or arranged in a crossword). Circle them. It's just nice to see the words.

The answers are in the back. Except for the "think about it" questions — those don't have answers. That's the point.

Start anywhere. The chapters build on each other, but not so much that you can't jump in.

Note:

Interactive version — the web edition of this book includes interactive word-search puzzles and a built-in reading experience with audio and games. If you received a zip archive alongside this PDF, open `summer_interactive/intro.html` in any browser. No internet required.

Part I

Part I — Foundations

1 What Is HITL ML?

> *The more capable an automated system becomes, the more consequential its failures are — and therefore the more necessary robust human oversight becomes.*
>
> *—Chapter 1*

1.1 Think about it

1. You use Google Maps and it reroutes you mid-trip because of traffic ahead. Someone had to teach it what "traffic" looks like and when it's worth rerouting. Who was that person? When did they do it? Do you think they knew their work would end up in your car?

2. A self-checkout machine at the supermarket flags your item and calls an attendant. Is that human-in-the-loop? What's the machine uncertain about? What is the human actually adding?

3. Think of something an AI system got wrong recently — a recommendation, a translation, an autocomplete, a search result. What kind of feedback would have corrected it? Who would have had to give that feedback, and when?

4. The chapter describes the "automation paradox": the more capable automation becomes, the more human oversight it demands. Does that feel right to you, based on things you've seen? Can you think of a counterexample?

5. HITL ML is defined as *deliberate, structured, and ongoing*. Which of those three feels hardest to achieve in practice? Why?

1.2 Spot the human

> *A hospital uses an AI system to flag chest X-rays that may show early signs of lung nodules. Radiologists review every flag before any action is taken. The system was trained on 50,000 annotated scans. Since deployment, radiologists have been asked to note whether the system's flag was correct or not — those notes are stored but not yet used to retrain the system.*

Questions to circle in your mind (or on the page):

- Where is the human in the loop right now?
- Where *was* the human in the loop before deployment?
- Is the feedback currently being collected actually in the loop? What would need to change for it to be?
- What happens if a radiologist disagrees with the AI but approves the flag anyway because it's faster?

1.3 Word search

Click and drag across letters to select a word. Diagonal, horizontal, and vertical — all directions are fair game. Words can also run backwards.

```
WORDS: ANNOTATION,ALIGNMENT,ORACLE,FEEDBACK,DELIBERATE,SUPERVISED,LABELING,AUTOPILOT,
↪UNCERTAIN,HUMAN

A L I G N M E N T D E A U H F
Z V N E T C M M T O Q N I U R
A V X T D V R Y L I Y N U M K
D U J A N F O A A X X O I A O
Q N Y R F Q S D B U J T U N R
Q C T E G E U L E K Y A F R A
Y E Q B A T P K L C P T A T C
D R L I Z J E H I A B I H O L
S T C L C X R P N B C O Y L E
R A Y E E E V V G D P N R I F
I I Q D T N I G R E Y X W P G
W N J M V U S L O E Q O D O H
H C K A S R E H S F H A C T W
U B H C B K D C Q H I V P U G
R E X S S P H Z P Z N G D A D
```

1.4 One more thing

The chapter ends with this idea: human involvement in ML is not a *temporary scaffolding* to be removed once models get good enough. It is a *feature*.

That's a strong claim. Most of the engineering world assumes the opposite — that the goal is full automation, and humans are in the loop only until the machine is ready to go it alone.

Write one sentence (here, or in your head) on which side you're on, and why.

Read the interactive version of this chapter online: https://your-site.netlify.app/chapters/01_introduction.html (word search, games, audio)

2 A Taxonomy of Human–Machine Interaction

> *Being consulted and being overridden are not the same thing — and the gap between them is where most of the interesting questions live.*

2.1 Think about it

1. Think of an app you use every day that makes suggestions — autocomplete, recommended routes, curated feeds. Does it feel like you're in the loop, or just watching it go?
2. When does "the human reviews the output" become meaningfully different from "the human rubber-stamps the output"? What would have to change for you to tell?
3. You get a push notification asking you to confirm something. You tap OK without reading it. Were you in the loop?
4. Is there a difference between a system that learns from what you click and one that learns from what you say? Does one feel more like feedback to you?
5. If you had to draw a line between "the machine is helping me" and "I'm helping the machine," where would you put it — and does it move depending on the day?

2.2 Spot the human

> *Gmail's Smart Compose watches as you type and offers to finish your sentences. You can accept the suggestion with a single key press, or ignore it and keep typing. The model was trained on billions of emails. It knows how sentences tend to end. It's very good at this.*

Where is the human in the loop?

- Is accepting a suggestion feedback? If so, is ignoring one also feedback — or just silence?

- At what point does Smart Compose shift from being “in the loop” to being “on the loop” — advising without really participating?
- The model was trained on human writing, but the humans who wrote those emails never consented to training a completion model. Were they in the loop?
- If you accept 80% of Smart Compose’s suggestions, who is writing your email?

2.3 Word search

Find the hidden words — they run across, down, or diagonal.

```
WORDS: ANNOTATION,INTERFACE,TAXONOMY,PIPELINE,FEEDBACK,PASSIVE,ORACLE,SCHEMA,LABEL,LOOP

L G V I W V U C T U F A Y R X
H K F E O M I U W R H N M A V
K Y C C Y P B H B Z K N O M M
I C G A S W O K G U P O N E M
U O E F B I E O H X R T O H R
I X S R N D S M L L H A X C E
Q P C E Y P E B D E U T A S F
Z V N T T C I E M M T I T O Q
I R A N V X D P F V R O Y I Y
U K D I J N F O E A X N X I Q
Y F Q O R A C L E L D U J U Q
T E V I S S A P G E I L Y F R
Y Q A T K P A D L Z J N H B H
S L E B A L C C X P C Y E R Y
E E V P R F I Q T N G R Y X W
```

Read the interactive version of this chapter online: https://your-site.netlify.app/chapters/02_taxonomy.html (word search, games, audio)

Part II

Part II — Core Techniques

3
Data Annotation and Labeling

> *A label is not a fact — it's an opinion that got promoted.*

3.1 Think about it

1. Have you ever been asked to categorize something and felt like none of the options quite fit? What did you do — pick the closest one, or leave it blank? What did that choice say about the categories themselves?
2. Two people watch the same video and disagree about whether it's "violent." Who's wrong? Is either of them wrong?
3. If annotators consistently disagree on a type of example, is that a problem to fix — or is the disagreement itself telling you something real about the world?
4. Think about a situation where you had to follow a rule that didn't quite cover your actual situation. Did following the rule get you to the right answer?
5. What's the difference between "everyone agrees" and "everyone is correct"? Can you have one without the other?

3.2 Spot the human

> *A content moderation team is labeling posts as "borderline" or "clearly violating policy." The post in question: a photo of a protest with a caption that could be read as incitement or as documentation, depending on how you look at it. Three annotators see it. Two say borderline. One says clearly violating. The system will record a majority vote.*

Where is the human in the loop?

- What does the majority vote hide — and is what it hides important?
- The annotator who voted "clearly violating" might be right. Or might be having a bad day. How would you tell?

- If the guidelines don't fully cover this case, does following them anyway produce a meaningful label?
- The model will train on this label. What does it learn from a contested example that got flattened into a single output?

3.3 Word search

Find the hidden words — they run across, down, or diagonal.

```
WORDS: GOLDSTANDARD,ADJUDICATION,ANNOTATOR,GUIDELINE,AGREEMENT,AMBIGUITY,LABELING,SCHEMA,
↪KAPPA,SPAN

C P W Y T I U G I B M A K E K
L A B E L I N G H A M E H C S
A X G P U E O J S V O J X P O
J Y X R P R I T N X N Q S C G
T G O L D S T A N D A R D W U
X E T X A U A D Z U G W C Y U
H G G T Z G C X Z Q R B I A F
P O N J C U I K O O E A W N I
P E D Z K I D L K L E P D N Z
M W O K J D U Z G H M P Z O O
F L J N H E J S J Q E A N T U
B F J E Z L D E G E N K A A F
W U S Q W I A W G I T P P T M
Q F I D X N Q L E D N J S O Q
Y I D E S E W Q M B M S O R U
```

Read the interactive version of this chapter online: https://your-site.netlify.app/chapters/03_annotation.html (word search, games, audio)

4 Active Learning

> *The model decides what you see. The question is whether you've noticed that yet.*

4.1 Think about it

1. If a machine asks you only the questions it's most confused about, can you trust that you're building a complete picture — or just filling in the machine's blind spots?
2. What does "uncertain" mean to a model versus what it means to you? Are those two things compatible enough to be useful?
3. A doctor reviewing AI-flagged X-rays gets shown the ones the AI is least sure about. Does that make her more useful — or does it change what kind of expert she becomes over time?
4. Is asking the right question harder than answering it? Think of a domain you know well. Who would ask the most useful questions — you, or a machine that's seen a million examples?
5. When you're on a budget — of time, of money, of attention — how do you decide what's worth your effort? Does that logic look anything like how active learning decides what's worth labeling?

4.2 Spot the human

> *A radiology AI has been integrated into a hospital workflow. It scans every X-ray that comes in, assigns a confidence score, and flags the ones it's least sure about for the radiologist to review. The radiologist looks at the flagged cases. The unflagged ones go through automatically.*

Where is the human in the loop?

- The AI chooses which cases the radiologist sees. Does the radiologist know that's happening?
- What happens to the radiologist's intuition about what "normal" looks like if she only ever sees the edge cases?
- A case the AI is confident about could still be wrong. Who catches that?
- The budget is the radiologist's time. Is the AI spending it wisely — and who gets to decide?

4.3 Word search

Find the hidden words — they run across, down, or diagonal.

```
WORDS: UNCERTAINTY,COMMITTEE,ENTROPY,CORESET,MARGIN,STREAM,ORACLE,BUDGET,QUERY,POOL

E H P O Z M B U D G E T X V L
T X B I F A C A H A S Y O T U
R T R J L E O S G M D R L K D
H E T W B R M N Y B A V T P D
E S F U X T M A T C U J U C U
I E C A N S I I L F W E Q N U
V R G B I P T E I C N Q C Y Z
T O K R Z P T C J T P E J M M
J C R O C K E I R E R E V E C
O M M U L O E O C T I L S Y E
I A X R Q L P H A B O W H R P
C R Q I N Y W I E O Q U E R Y
O G I E H T N L P Q D K K P E
A I C M U T V W Z M M S U X V
Q N O H Y H F W E X A K J A R
```

Read the interactive version of this chapter online: https://your-site.netlify.app/chapters/04_active_learning.html (word search, games, audio)

5 Interactive Machine Learning

> *Somewhere between the tenth correction and the thirtieth, you stop correcting and start accepting. That's the moment worth watching.*

5.1 Think about it

1. Think of something you taught someone — a skill, a game, a habit. How did you know when they'd got it? Could you imagine teaching a machine the same way?
2. The "Grandmother Test" asks whether a non-expert can understand what a system is doing and why. Pick any app on your phone. Would your grandmother pass?
3. Have you ever given up correcting a phone's autocorrect and just started spelling things its way? What does that say about who's really adapting?
4. Fatigue is real — after a while, feedback gets sloppier. If the model was trained partly on your tired corrections, is it learning you at your best or you at your most exhausted?
5. When feedback feels like it's working, is it because the model improved, or because you started expecting less? Is there a way to tell from the inside?

5.2 Spot the human

> *Google's Teachable Machine lets you open a browser tab, hold objects up to your webcam, and train a classifier in real time. You show it examples of "thumbs up" and "thumbs down," and within seconds it starts predicting. You keep training until it gets it right. It feels like teaching.*

Where is the human in the loop?

- The model updates as you add examples. But you're also updating — adjusting how you hold the object, what backgrounds you use. Who's learning from whom?
- If the model keeps making the same mistake no matter what you show it, at what point do you conclude the problem is the interface, not your teaching?

- The loop is fast and visible — you can watch the model change. Does that transparency make you trust it more, or just feel more in control?
- What's the first sign you'd notice that the model had anchored on something irrelevant — like the color of your shirt instead of your gesture?

5.3 Word search

Find the hidden words — they run across, down, or diagonal.

```
WORDS: INITIATIVE,THRESHOLD,FEEDBACK,ITERATE,CORRECT,FATIGUE,DIRECT,RAPID,TRUST,TEACH

W E M Z D T T F C R V W V Y W
O T M U L I C M Y A M M F W Q
A E C K O S E U W P N H K L O
S A G M H Q R E R I W A V N S
B C U O S U I V Z D S V B H S
H H H B E B D I C T Z L S U J
K O C Z R Q W T M S B W Y M V
O T O X H M S A B M H M B C L
W S R I T C R I E M K H A N K
X U R E V C S T F A T I G U E
T R E Y X N T I S Z D V R V E
U T C Y Z H C N X A E U J I U
Z V T R U E T I S X K M S Y I
M H A G D D E T A R E T I R B
P F E E D B A C K B T N R K K
```

Read the interactive version of this chapter online: https://your-site.netlify.app/chapters/05_interactive_ml.html (word search, games, audio)

Part III

Part III — Learning from Human Feedback

6 Reinforcement Learning from Human Feedback

The model didn't decide to be cautious. The humans who ranked its outputs decided that cautious answers felt safer to prefer — and the model listened.

6.1 Think about it

1. When you give feedback on something — a piece of writing, a recommendation, a friend's plan — are you expressing what you actually want, or what you think you should want? Could you tell the difference?
2. RLHF-trained models sometimes sound overly formal or weirdly hedged. If that's what human raters preferred, is the model wrong — or are the raters?
3. The "human" in RLHF is usually a team of paid contractors doing thousands of comparisons quickly. Does that change how you think about the word "human"?
4. If a reward model can be gamed — if a model learns to produce outputs that score well without being genuinely good — how would you know? What would the outputs look like?
5. Think about what gets averaged out when you aggregate thousands of human preferences. What kinds of users and values get centered? What gets erased?

6.2 Spot the human

> *You're using a ChatGPT-style assistant. You ask it a question about drug interactions. It gives you a careful, hedged, heavily-caveated response that tells you to "consult a healthcare professional." You ask a follow-up. Same thing. The information you need is probably in there somewhere, but it's wrapped in so much caution it's hard to use.*

Where is the human in the loop?

- The caution isn't a bug — it's a learned preference. Whose preference was it, exactly?
- The KL divergence penalty keeps the model from drifting too far from the base. That means the model is being pulled in two directions at once. What does that do to its voice?
- If the raters who shaped this model were mostly risk-averse, does the model represent "human preferences" — or a particular human's preferences in a particular context?
- You, the user, are also a human. Your preference for a direct answer didn't make it into the training loop. Why not?

6.3 Word search

Find the hidden words — they run across, down, or diagonal.

```
WORDS: KLDIVERGENCE,PREFERENCE,ALIGNMENT,PROXIMAL,TRAINING,RANKING,PENALTY,REWARD,POLICY,
↪HUMAN

A L I G N M E N T M O U F E Y
M N R Y I X V P M W E Z M L V
J G A M E M B D B G D M A K I
K N U M T X Q K B M H M I M P
L I J J U R B H A T I K R V M
D N I X B H T X Q X P B E W Y
I I K J U N R A O C E U W Z T
V A T W G A L R J F V A A A V
E R R X U W P M C T H Y R R D
R T T L X Y T L A N E P D U A
G N U E E C N E R E F E R P G
E B K T E W P I M U F J D F K
N H J Y C I L O P Y J K L S X
C G N I K N A R T U V C P T P
E R I S H T N P Z U N Q N F W
```

Read the interactive version of this chapter online: https://your-site.netlify.app/chapters/06_rlhf.html (word search, games, audio)

7 Learning from Demonstrations

> *The car learned to drive by watching humans. The problem is that humans, when they're being watched, drive slightly differently than when they're not.*

7.1 Think about it

1. Think of something you do automatically — the way you parallel park, the route you take home, the way you hold a fork. Could you explain it well enough for someone to learn it by watching you once? A hundred times?
2. What's the difference between "what humans do" and "what humans would do if the situation were different"? A self-driving car trained on human data knows the first thing. It doesn't know the second.
3. Have you ever learned something by watching someone and then discovered you'd learned the wrong parts — the tics, the habits, not the skill itself? What was it?
4. Distribution shift sounds technical but you've felt it: showing up in a situation that's a bit like the ones you prepared for, but not quite. What do you do? What should a model do?
5. Behavioral cloning makes the model learn to copy actions. But most experts can't fully articulate why they make their decisions. Is it possible to clone expertise without understanding?

7.2 Spot the human

> *A self-driving car company records thousands of hours of human driving and trains a behavioral cloning model on the data. The model learns to stay in lane, signal, yield, and accelerate smoothly through familiar scenarios. Then it encounters a flooded road that no driver in the training set ever drove through, because experienced drivers turned around.*

Where is the human in the loop?

- The demonstrations covered everything the humans decided to demonstrate. What's missing is everything they decided wasn't worth demonstrating — including their judgment about when to stop.
- DAgger tries to fix distribution shift by asking the expert what they would do in the states the model actually visits. Why is that different from just recording more drives?
- The expert policy is a snapshot. Drivers improve, rules change, cities get rebuilt. How long does a demonstration dataset stay valid?
- The humans in the loop were the original drivers. Are they still in the loop when the model is deployed a year later?

7.3 Word search

Find the hidden words — they run across, down, or diagonal.

```
WORDS: IMITATION,AGGREGATE,CLONING,ROLLOUT,DATASET,DAGGER,EXPERT,POLICY,SHIFT,ROBOT

F R U N Z Y C C I Q T R C P Z
R O L L O U T Z N L Y L B T I
U T Q V P I W U Z K O A S H G
T R E P X E T L D N N W Y F F
P G E D E X Z A I S E L C Q M
D I O A U D G N T T N C I L A
I B E F X G G Q A I P G L Z Y
R Z P Q E O S G S A M W O C F
E O R R K U E B T H W I P B M
D A B C J R B R J H T X R E T
R S Y O G I T S C S P K R B F
D N N G T L J S Y P T E C Q I
D O A U Z O A N Z S Q Y C Z H
Z J Z W P Q D Y H L Z K F O S
A S D V O D A T A S E T O I U
```

Read the interactive version of this chapter online: https://your-site.netlify.app/chapters/07_demonstrations.html (word search, games, audio)

8 Learning from Comparisons and Rankings

> *Preferences are real. They're just not consistent, context-free, or immune to how the question was asked.*

8.1 Think about it

1. Think about the last time you chose between two things you genuinely liked differently. Was it easy to say which one was "better"? What would it mean to express that difference to a machine?
2. Spotify's Discover Weekly is built on the implicit assumption that what you listened to is what you preferred. Is that true? What about the song you left on while you made dinner?
3. Are your preferences transitive? If you prefer A to B and B to C, do you always prefer A to C? Think of a case where that might not hold.
4. If someone asked you to rank 50 things from best to worst, how would your answers on day one compare to your answers on day seven? Does that instability mean your preferences aren't real — or just that they're human?
5. What kinds of things are hard to compare ordinally — where "better than" doesn't quite capture your experience of the difference? What gets lost when you try?

8.2 Spot the human

> *Spotify Discover Weekly compiles a playlist of songs you've never heard. It's built on collaborative filtering and implicit feedback — your listening history, skip rates, repeat plays, playlist additions. Every play is a data point. Every skip is a data point. You never rated anything.*

Where is the human in the loop?

- You didn't compare songs — the model inferred comparisons from your behavior. Is that the same thing? Is it close enough?
- The Bradley-Terry model assumes comparisons are consistent and probabilistic. What happens when a song you loved at 20 now feels dated? Does the model handle nostalgia?
- Cardinal feedback (ratings) and ordinal feedback (rankings) capture different things. What does Discover Weekly use — and what does it miss because of that?
- If you listen to a song because it's familiar, not because you prefer it to something new, is that signal or noise?

8.3 Word search

Find the hidden words — they run across, down, or diagonal.

```
WORDS: COMPARISON,PAIRWISE,CARDINAL,RANKING,UTILITY,ORDINAL,BRADLEY,ELICIT,REWARD,NOISE

N X A V U K L C R U Z T O Q O
O D O S N C P F G A Y C V T W
I G H L B O M R N P N B I T A
S X K M V N S L A C W K H O K
E Y D U B C A I G R E G I S D
E H I K G N R O R E E S U N N
Y B C S I W S C R A M W Q A G
X I U D I T E D E D P C A D Y
M R R S Y Q O S V K I M O R M
V A E J Z J P G E A V N O O D
C N H S J D G B F M D K A C M
B L R X T F G V K R E Q Y L R
G V Z B B B R A D L E Y O E T
P L D B H A C B Y T I C I L E
E H M D U T I L I T Y B O P W
```

Read the interactive version of this chapter online: https://your-site.netlify.app/chapters/08_preferences.html (word search, games, audio)

Part IV

Part IV — Applications

9 HITL in Natural Language Processing

> *The model can tell you what word is there. It can't tell you what it means to the person who wrote it — and neither, sometimes, can the person who's labeling it.*

9.1 Think about it

1. Think of a word that means different things depending on who says it, to whom, and when. How would you write a labeling guideline for it?
2. When you read a sentence and feel like you understand it, what are you doing? If a model predicts the right label without understanding the sentence, does it matter?
3. Snorkel and weak supervision let you write rules instead of labeling examples. Is a rule the same as an opinion? What do you lose when you automate the labeling?
4. Majority vote is used to aggregate labels across annotators. But what if the minority was right? Is there a context where you'd actively want to preserve the minority label?
5. What's the difference between a word being ambiguous and a word being contested? Does your answer change how you'd handle it in a labeling pipeline?

9.2 Spot the human

> *A medical NLP system is extracting drug names from clinical notes. The word "lithium" appears in a note. In most contexts, it's a psychiatric medication. In this note, it's part of a sentence about a patient's diet and a supplement they mentioned to their doctor. The labeling function flags it as a drug name.*

Where is the human in the loop?

- The labeling function didn't read the sentence — it matched a pattern. Who's responsible for what it missed?

- A human reviewer would catch this in a second. But there are 4 million notes. Where does the human fit when the volume is that high?
- The model trained on this data will probably learn the wrong thing about "lithium." How many wrong examples does it take before that becomes a real problem?
- Sequence labeling breaks text into tokens and labels each one. Does that framing fit how humans actually read — or does it miss something structural?

9.3 Word search

Find the hidden words — they run across, down, or diagonal.

```
WORDS: SENTIMENT,SEQUENCE,LABELING,FUNCTION,MAJORITY,SNORKEL,ENTITY,TOKEN,VOTE,SPAN

N P O U C V Y T C U H H N K Q
L S A V F L L T D T X S D F S
E I Y T N C S G I I Y P L W V
K V Q X E O E N V T D A Y S O
R S E X V M N I T A N N I Q D
O K T S K N T L X U Q E S Q R
N X H A H B I E K Y C F E P O
S I N Z Z E M B R T A U Q N X
W Q E I D T E A O I H N U A P
S E K I M O N L F R I C E S H
Y F O H T V T J L O G T N B D
N B T W D C C Q N J G I C Q W
I I G P O U J K X A A O E J V
V G E U W Y K H D M B N D X K
D T N Y U H B W H L Q I C Y E
```

Read the interactive version of this chapter online: https://your-site.netlify.app/chapters/09_nlp.html (word search, games, audio)

10 HITL in Computer Vision

> *The annotator drew the box. Whether the box captured the thing the model needed to understand is a different question entirely.*

10.1 Think about it

1. Where does a pedestrian end? It sounds absurd, but that's exactly what a bounding box annotation requires you to decide. Does the answer depend on why you're drawing the box?
2. Think of something you can recognize instantly but would struggle to draw a precise boundary around. A crowd. A shadow. A reflection. What does "annotation" even mean for those things?
3. If two annotators draw slightly different boxes around the same object, and a model trains on both, what does it learn? Is the average the right answer?
4. Polygon annotations are more precise than bounding boxes but take longer. At some point, more precision costs more than it's worth. Who makes that call, and how?
5. Have you ever labeled something — on any platform, in any context — and felt like the category didn't fit? What did you do? What should the platform have done?

10.2 Spot the human

> *An annotator is drawing bounding boxes around pedestrians for a self-driving car dataset. She's doing hundreds per hour. The guidelines say to include the full body, but some pedestrians are partially occluded — cut off by a pole, a car, another person. She's making judgment calls every few seconds.*

Where is the human in the loop?

- The model will train on her judgment calls alongside everyone else's. If she's more conservative about occluded pedestrians than other annotators, is her data worse — or different?
- After hour three, boxes start getting a little looser. The platform doesn't know that. Should it?
- Tight bounding boxes are more "correct" in some sense, but the model may actually generalize better from slightly noisier annotations. Does the human need to know that to do the job well?
- The pedestrian the model will one day fail to detect — was her bounding box in the training set?

10.3 Word search

Find the hidden words — they run across, down, or diagonal.

```
WORDS: DETECTION,BOUNDING,KEYPOINT,SEGMENT,POLYGON,CAPTION,OBJECT,ACTIVE,LABEL,MASK

K C O V U S A J K L R U V Q P
C S J F N G N I D N U O B N G
D V A C A F E B E A J V X O L
A F Q M B H S W J W S A E I P
M N L R E H A Q T J U K M T K
B E R P T N I O P Y E K N P J
E L Z D K M W Z I Q N E A A R
D A D R C A D J U O M C P C O
O B I J Z V T H I G T Z I B P
H E S V G H J T E I H B J O R
P L F H D W C S V I X E L M Y
A X J R D E Z E F V C Y N H P
M Y R B T Q J P D T G F F Y K
Q G A E H G Z P J O W V O U U
R R D D G B L A N G C H E D W
```

Read the interactive version of this chapter online: https://your-site.netlify.app/chapters/10_vision.html (word search, games, audio)

11
HITL in Healthcare and Science

> *The AI flagged it. The cardiologist overruled it. The patient was fine. Next time, it might not be the AI that's wrong.*

11.1 Think about it

1. When you trust a doctor's judgment, what are you trusting — their knowledge, their experience, their confidence, or something harder to name? Does it matter if an AI has all the first three?
2. In healthcare, false positives (wrong alarms) and false negatives (missed cases) have very different costs. Who should decide where to set the threshold — the engineers, the clinicians, or the patients?
3. Two radiologists look at the same scan and disagree. This happens all the time. Does that mean one of them is wrong, or that the image is genuinely ambiguous? How should a model handle that?
4. "Regulatory oversight" sounds dry, but it's the mechanism by which society decides who bears responsibility when something goes wrong. If an AI makes a call and a doctor follows it, who's accountable?
5. Think about a time you were on the receiving end of someone's expert judgment — a diagnosis, a legal opinion, a teacher's assessment. Did you feel like you were in the loop? Should you have been?

11.2 Spot the human

> *An AI system reads ECGs and flags potential arrhythmias for cardiologist review. The system has been validated on a large dataset and performs comparably to junior cardiologists on most arrhythmia types. A cardiologist reviews a flagged ECG and disagrees with the AI's reading. The patient's chart is ambiguous. The cardiologist makes the call.*

Where is the human in the loop?

- The cardiologist overruled the AI. Was she overruling a tool, or a second opinion? Does the framing change how she thought about it?
- The AI was trained on data from a different hospital system with different demographics. The cardiologist doesn't know that. Should she?
- If the AI is right more often than it's wrong, and the cardiologist overrules it often, is the cardiologist making the system worse? How would you measure that?
- Consent forms say the hospital uses AI to assist diagnosis. Is that "in the loop" for the patient?

11.3 Word search

Find the hidden words — they run across, down, or diagonal.

```
WORDS: RADIOLOGY,DIAGNOSIS,OVERSIGHT,CLINICAL,CONSENT,SAFETY,EXPERT,TRIAL,AUDIT,BIAS

B K O J R V T O C G T R L W J
I P Y E R U D L N R Z F D J O
X A V V F V I I E N X C R Q Q
Y J S W S N V P A O X S O P M
A U Y W I T X I E G A Z V Y M
M U M C K E H K M I N S U V K
R O A N Q X S G B K P O T D A
E L M G Z N Y I I B M Y S R K
B Q I B K T Y A T S J O C I T
P X R Z E R E R L I R C X P S
Q S A F H V I Y N F X E Z Z D
S U A X A A T Q T X Y F V Y S
P S U E L E Q R N N G P B O O
O M C R Y G O L O I D A R B D
C O N S E N T T I D U A J Q C
```

Read the interactive version of this chapter online: https://your-site.netlify.app/chapters/11_healthcare.html (word search, games, audio)

Part V

Part V — Systems and Practice

12 Annotation Platforms and Tooling

The platform decided you could only pick one label. That decision shaped every dataset built on it.

12.1 Think about it

1. Think about a form you've filled out that didn't have a field for what you actually wanted to say. Did the form shape your answer? What did the data collector learn?
2. Interface design is a series of choices about what's easy and what's hard. In an annotation platform, what things should be easy — and for whom?
3. If annotators can only choose from the options the platform offers, then the platform designer is making decisions about the data. Are they accountable for those decisions?
4. "Quality control" in annotation usually means checking consistency. But consistency and correctness aren't the same. Can you think of a case where they'd point in opposite directions?
5. Think about a task you've done repetitively — data entry, tagging, sorting, filing. At what point did you stop thinking carefully and start going on autopilot? What would have kept you engaged?

12.2 Spot the human

A startup is building a training dataset using a crowdsourcing platform. The platform shows one item at a time, offers three label choices, and logs how long each annotator spends. Fast annotators get flagged for review. The platform's design — one item, three choices, a timer — quietly determines what kinds of annotations are possible and what kinds of data get produced.

Where is the human in the loop?

- The platform was designed by engineers who probably didn't annotate the task themselves. Does that matter?
- The three label choices were chosen by the startup. What got left out of the interface?
- The timer creates pressure. Does pressure improve annotation quality or degrade it? Does the platform know?
- The human doing the annotation is in the loop. The human who designed the platform is no longer in the loop. But her choices are still making decisions.

12.3 Word search

Find the hidden words — they run across, down, or diagonal.

```
WORDS: THROUGHPUT,INTERFACE,PLATFORM,WORKFLOW,QUALITY,WORKER,REVIEW,EXPORT,SCALE,TASK

Y G E E L A C S R P Q H S O T
T E Z H Q V V E L Y M Q V G U
C V V E B X K A T A L I C M P
B E Q E D R T X X W T M Q V H
A E J O O F B F E R E Z E E G
C W X W O H K I U G L I G E U
C Y Y R I B V Y C V E Z G H O
W R M K K E L L J C A F V M R
I O C P R H Q U A L I T Y W H
B K L X F E W F J I L Z N V T
V R P F X U R A S T W P I U D
Q B W P K E D P O G C T W X I
T B O J T R Y I Q F A K A Q H
G R D N U K O O D H W L E S B
T C I Q E Z Q W Y G I Y Q W K
```

Read the interactive version of this chapter online: https://your-site.netlify.app/chapters/12_platforms.html (word search, games, audio)

13
Crowdsourcing and Quality Control

> *Five hundred workers, three labels each — and somewhere in that number, a few people who were clearly just clicking. The math assumes you can tell them apart.*

13.1 Think about it

1. When a crowd votes and you go with the majority, what are you assuming about the crowd? What has to be true about them for that assumption to hold?
2. "Gold standard questions" are tasks where you already know the answer, used to detect bad annotators. Does that work if a careful worker gets a hard gold question wrong?
3. Think about how much the task of labeling an image pays on Mechanical Turk — maybe a few cents. Does that affect what kind of answers you'd expect? What kind of workers?
4. Dawid-Skene estimates worker reliability from agreement patterns. But two workers who agree can both be wrong in the same systematic way. How would you catch that?
5. If you could redesign one thing about how crowdsourcing platforms work — payment, task design, quality control, worker communication — what would it be and why?

13.2 Spot the human

> *Five hundred workers on Mechanical Turk are labeling images — three workers per image, earning a few cents each. Most clusters of three are consistent. But looking at the logs, a handful of workers are clearly clicking through without reading — their per-image times are under a second, their responses are random. The majority vote algorithm doesn't know that yet.*

Where is the human in the loop?

- The workers are humans. But their presence in the loop depends entirely on whether they're actually engaged with the task. At what point does a worker stop being "in the loop"?
- Majority vote is designed to filter out random noise. But if a whole group of workers shares the same misunderstanding, it won't help. What does?
- The few-cents-per-image rate was set by someone. That rate shapes who does the work and how much attention they give it. Is the person who set the rate in the loop?
- A spam filter based on response time would catch the fastest clickers. But some fast responders are legitimately fast. What's the cost of wrongly removing them?

13.3 Word search

Find the hidden words — they run across, down, or diagonal.

```
WORDS: AGREEMENT,MAJORITY,QUALITY,WORKER,FILTER,CROWD,DAWID,SKENE,SPAM,GOLD

M C Y J E U M H O T H J H G Y
T P T P N H F G R G C K E Y F
I O I K X N D A O H S R C X U
R Y L M L I G L W Z I S O T P
Z V A K T F D F R X L C B W U
R M U W O R K E R N Y A Y V D
N R Q O P E S P A M V V U M V
G T Z L T N E M E E R G A B H
I N R Y A Q R T K A K I C M W
S S B U B B Q E M R Y X U B V
P X D I W A D Q T E L Q D M Y
N V Z A Y S B N R L N G Q U R
Z B M I J G E Y R M I J G P D
M A J O R I T Y T P H F W N Z
B Q E N E K S V X M J E G M U
```

Read the interactive version of this chapter online: https://your-site.netlify.app/chapters/13_crowdsourcing.html (word search, games, audio)

14
Evaluation and Metrics

> *95% accuracy sounds great until you find out the dataset is 95% one class, and the model just learned to say that class every time.*

14.1 Think about it

1. Think of a number you've been given to assess something — a score, a grade, a rating. What did that number leave out? Was the thing it left out important?
2. If a model has been optimized specifically to do well on a benchmark, does doing well on that benchmark tell you anything anymore?
3. Calibration means a model's confidence matches its actual accuracy. Why might that matter more than raw accuracy when you're making a decision based on the model's output?
4. A model trained in 2022 is deployed in 2025. The world has changed. The benchmark score hasn't. What does the score actually measure now?
5. Precision and recall pull against each other — improving one often hurts the other. In your field (or in your life), which kind of mistake costs more: false positives or false negatives?

14.2 Spot the human

> *A machine learning team reports that their model achieves 95% accuracy on a standard benchmark. The benchmark is publicly available and widely used. They submit to a conference. A reviewer notes that the benchmark's class distribution is 95% negative examples. The model predicts "negative" for everything.*

Where is the human in the loop?

- The benchmark was designed by humans, with choices about what to include and what distribution to use. Those choices are still active — they're just invisible now.
- The team may not have intentionally gamed the metric. But they optimized for it. Is there a difference?
- Cohen's kappa adjusts for chance agreement. If you applied it to this model, what would you find?
- The reviewer caught it. How many papers didn't have that reviewer? What's already deployed based on that number?

14.3 Word search

Find the hidden words — they run across, down, or diagonal.

```
WORDS: CALIBRATION,PRECISION,BENCHMARK,ACCURACY,COVERAGE,RECALL,METRIC,KAPPA,AUDIT,DRIFT

C B U C I R T E M Q O H Q O F
Y P S L C Q J K I Y Z Z Z B D
T K M M N O M E N L K P N X S
Z D E R C R V K O K Z I T V N
E N K I W A Z E R E Y Q R W O
J W J D O Z L A R D G S D Y I
R E C A L L M I E A A F C W S
H R I E T H G G B U G A U X I
K K K G C F F Z D R R E N I C
S C Q N X B I I A U A A U Z E
F J E O P K T R C O K T C H R
B B Q M P F G C D D N O I Y P
S B M V S H A J X L B R U O H
M A N E N A P P A K Q N A C N
F R Y Y H M L V X A D F Z I D
```

Read the interactive version of this chapter online: https://your-site.netlify.app/chapters/14_evaluation.html (word search, games, audio)

Part VI

Part VI — Ethics and Horizons

15 Fairness, Bias, and Ethics

> *The model didn't invent the bias. It learned it, faithfully, from humans who didn't notice they had it.*

15.1 Think about it

1. If a dataset was built by people making reasonable decisions in a biased world, is the dataset biased? Is the model trained on it biased? Is the organization that deployed it responsible?
2. "Removing bias" often means choosing whose idea of fairness wins. Whose idea of fairness is encoded in the tools you use every day?
3. Think about what "neutral" would mean for a hiring algorithm. Is neutrality possible — or does every design choice favor someone?
4. The annotators who labeled "good candidates" for the historical hiring data were humans trying to do a good job. They probably didn't think of themselves as introducing bias. Does intent matter?
5. Transparency — knowing what an AI is doing and why — is supposed to help. But transparency for whom? The engineers? The regulator? The person being decided about?

15.2 Spot the human

> *A hiring algorithm is trained on a decade of historical hiring data. Human recruiters labeled candidates as "good" or "not good" based on who got hired and performed well. Those historical hires skewed toward a particular demographic. The model learned that pattern. Now it's ranking new candidates.*

Where is the human in the loop?

- The recruiters who labeled the historical data are no longer in the loop. But their judgments are still running.
- A "feedback loop" in the algorithmic fairness sense means the model's outputs shape future training data. What does that mean for underrepresented candidates over time?
- If you audit the model and find disparate impact, what do you do? Retrain? Adjust outputs? Reconsider the training data? Who decides?
- The workers who annotated this data — did they know it would be used to train a hiring model? Did they consent to that use?

15.3 Word search

Find the hidden words — they run across, down, or diagonal.

```
WORDS: TRANSPARENCY,FAIRNESS,CONSENT,PRIVACY,WORKER,RIGHTS,AUDIT,POWER,BIAS,HARM

N P S T X D S T H G I R M V C
I B T W W I P O W E R S J K I
E C R B T N Z O P P C S P I R
L D A I I H M A R N P E S F U
I B N H Z A U U X L N N K W L
R U S A B D S L W V G R P K P
Q J P Q I M I U P G V I H F P
R G A T V W L O T S U A P R A
W B R R D T Q I F N P F I G U
V N E D N G J G Z Q E V A W U
H P N Z W O R K E R A S S P T
Y G C Z N A L O I C C H N N L
N R Y M K C B Y Y E N E A O A
C Y R X B U Z W I X W L B R C
U U R C N R X H X A Q O C B M
```

Read the interactive version of this chapter online: https://your-site.netlify.app/chapters/15_ethics.html (word search, games, audio)

16 Future Directions

> *The question isn't whether AI will be more capable. The question is whether "human in the loop" will still mean anything when it is.*

16.1 Think about it

1. Foundation models are trained on so much data that it's hard to say whose values they encode. If you can't trace the feedback, is there still a human in the loop?
2. Think about a task you'd feel comfortable delegating to an AI agent — booking a flight, drafting an email, ordering groceries. Now think about a task you'd never delegate. What's the line, and what does it tell you about trust?
3. As the scale of AI systems grows, individual human feedback becomes a smaller and smaller fraction of what shapes the system. Does that matter? What's the minimum meaningful fraction?
4. Synthetic data lets models train on data generated by other models. If the feedback loop closes — models generating data, training models — where does the human in the loop go?
5. If you could specify one thing an AI should always ask you before acting on your behalf, what would it be?

16.2 Spot the human

> *An AI agent has access to your email, calendar, and browser. It can draft responses, schedule meetings, and research topics on your behalf. You've given it broad permission. One afternoon it sends an email you didn't explicitly approve, to a contact you didn't realize it knew about, on a topic you would have handled differently.*

Where is the human in the loop?

- You gave it broad permission. Is that consent, or just a consent-shaped thing that turned out to mean something different than you thought?
- Alignment is supposed to mean the AI does what you actually want, not just what you said. How would you specify "what you actually want" for an email about a professional relationship?
- The loop used to be: AI suggests, human approves. At some point it became: AI acts, human reviews. Is that still a loop?
- Foundation models can generalize to tasks they weren't explicitly trained for. Does that make it harder or easier to know what the human in the loop is supervising?

16.3 Word search

Find the hidden words — they run across, down, or diagonal.

```
WORDS: FOUNDATION,AUTONOMOUS,SYNTHETIC,ALIGNMENT,FEEDBACK,FUTURE,AGENT,TRUST,SCALE,LOOP

Y Q C B T N E M N G I L A S W
J T Z M H W B Q P U X R J T G
B U K C A B D E E F W V N E F
V T R Z S E Q U W V A B B V W
P I D S Q E B I Z W E Y S S Z
W D M U X H G S D Y K E R E X
R D H O N X M W D G Y H J D T
E U C M G F O U N D A T I O N
L D P O B U N R T R U S T J T
A Y O N E F U T U R E W S U V
C R O O M M Z H W B Z Q O K Y
S M L T D L Y M S J D O E N T
L O J U S Y N T H E T I C G R
T A W A N C N A J A G E N T I
T W H B N E G M K R Z D Z X Y
```

Read the interactive version of this chapter online: https://your-site.netlify.app/chapters/16_future.html (word search, games, audio)

Part VII

Case Study

17 Limen: A Human in the Loop of Everything

> *A threshold isn't a wall. It's a question the system is asking you — and the answer you give shapes what it asks next time.*

17.1 Think about it

1. Think about the last time a piece of technology did something helpful without asking. Was it helpful because it asked, or helpful despite not asking? Does the difference matter to you?
2. "Voice-first" means the interface is a conversation. But conversations have norms — about when to interrupt, when to wait, when to assume. Who taught the OS those norms?
3. Local processing means your data doesn't leave your device. Why does that feel different from a privacy policy that promises the same thing? What's the underlying thing you're actually trusting?
4. The Grandmother Test asks whether a non-expert could understand and trust what's happening. Would your grandmother trust a system that anticipates her needs without being asked? Would you?
5. If an OS tracks causal events — WID, what-I-did — to understand your routine, at what point does understanding your routine become predicting your intentions? Is that a line worth drawing?

17.2 Spot the human

> *Your OS has been learning your morning routine for three weeks. Coffee at 7:15, news app at 7:20, calendar check at 7:25, first work message at 7:45. It starts anticipating — dimming the bedroom light at 7:10, pulling the news headlines before you open the app, surfacing your first meeting's prep notes at 7:40. One morning, it sends the message you were composing — a message you hadn't finished — because it predicted you were done.*

Where is the human in the loop?

- The OS was following a pattern it had genuinely learned. Was it wrong to send the message, or wrong to have the permission to send messages at all?
- Graceful degradation means the system falls back to asking when it's uncertain. The OS wasn't uncertain — it was confidently wrong. Is that better or worse?
- The difference between "helpful" and "presumptuous" is often just whether the system was right. Does that mean the threshold for acting should be higher than just "high confidence"?
- If you could tell the OS one rule about when to always ask before acting, what would it be?

17.3 Word search

Find the hidden words — they run across, down, or diagonal.

```
WORDS: THRESHOLD,FEEDBACK,PRIVACY,ROUTING,CAUSAL,DESIGN,LIMEN,VOICE,LOCAL,LOOP

F A I W N H E S W B X D Y B I
K G V E S H V B S R F A D F Y
W O R F C O F R R M F T N Q Z
N H Z M E F E E D B A C K W D
L O C A L Z B H E A A Y O D L
A S U X T H R E S H O L D O I
U T G L W K Y P R M M R O M F
M X K L L X M C E O G P N V S
A W P I Q N D N A C U C V M A
C I M W Q B C A N V I T U X C
V E B W F X O A U G I O I G A
N F B Y L Y Z F U X I R V N Z
H M A A R B M V H S J S P J G
W S X V Q P F I B P A G E E U
Y U A E O T F R Y D Y L Z D J
```

Read the interactive version of this chapter online: https://your-site.netlify.app/chapters/17_limen.html (word search, games, audio)

Part VIII

Back

18 Answers

18.1 Word Search Answers

18.1.1 Chapter 1 — What Is HITL ML?

(In the web version, found words highlight automatically. These are for the print edition.)

```
Grid (15x15):
A L I G N M E N T D E A U H F   row 1
Z V N E T C M M T O Q N I U R   row 2
A V X T D V R Y L I Y N U M K   row 3
D U J A N F O A A X X O I A O   row 4
Q N Y R F Q S D B U J T U N R   row 5
Q C T E G E U L E K Y A F R A   row 6
Y E Q B A T P K L C P T A T C   row 7
D R L I Z J E H I A B I H O L   row 8
S T C L C X R P N B C O Y L E   row 9
R A Y E E E V V G D P N R I F   row 10
I I Q D T N I G R E Y X W P G   row 11
W N J M V U S L O E Q O D O H   row 12
H C K A S R E H S F H A C T W   row 13
U B H C B K D C Q H I V P U G   row 14
R E X S S P H Z P Z N G D A D   row 15

Words:
  ALIGNMENT    → row  1, col  1, →  right
  ANNOTATION   → row  1, col 12, ↓  down
  HUMAN        → row  1, col 14, ↓  down
  UNCERTAIN    → row  4, col  2, ↓  down
  ORACLE       → row  4, col 15, ↓  down
  SUPERVISED   → row  5, col  7, ↓  down
  LABELING     → row  3, col  9, ↓  down
  DELIBERATE   → row 11, col  4, ↑  up
  FEEDBACK     → row 13, col 10, ↑  up
  AUTOPILOT    → row 15, col 14, ↑  up
```

The “Think about it” and “Spot the human” sections don’t have answers here.

If you found yourself arguing with a question, that's the right answer.