African Institute for Mathematical Sciences — Cameroon
Structured Master's, Data Science Stream · Computer Vision

Vision You Can Trust

Week 3 · Day 4: Robustness, Fairness, Ethics, and the Road Ahead

Yusuf Brima, Ph.D.

Introduction to Computer Vision · AIMS Cameroon, 2026–2027

Today

1. **When images change**: distribution shift, measured on real leaves
2. **Adversarial examples**
3. **Can we trust the confidence?** Calibration and uncertainty
4. **Fairness**: who does the model fail?
5. **Privacy, surveillance, consent and deepfakes**
6. **Responsible practice**: documentation, regulation, deployment
7. **The road ahead**: open problems and your place in them

By the end of today you should be able to

- test a model under realistic corruptions and shifts;
- read a reliability diagram and explain calibration;
- break an aggregate metric down by subgroup;
- name the main ethical risks of vision systems and concrete mitigations;
- write a short model card for your capstone.

Why a whole lecture on this?

- Every number in this course was measured on a test set drawn like the training set. Deployment never is.
- Vision systems now make or inform decisions about health, money, movement and policing.
- The people most affected by errors are often not the people who built the system, and are rarely in its training data.

Intuition

A model is a promise: "on images like these, I am right this often". Today is about checking the small print: which images, how often, for whom, and what happens when I am wrong.

Practitioner's Note

Your capstone rubric rewards an honest limitations section (Week 2, Day 5). Today gives you the tools to write it.

Part 1

When images change: distribution shift

The i.i.d. assumption, and why it breaks

- Training and testing assume images are drawn **independently** from the **same distribution**.
- Real deployments break this constantly:
  - new camera or phone;
  - new hospital, farm, city, country;
  - new season, time of day, weather;
  - the world itself changes (new vehicle models, new diseases).

Kinds of shift

Covariate shift: images look different, labels mean the same (a darker camera).

Label shift: class frequencies change (a disease outbreak).

Concept shift: the meaning of a label changes (a new definition of "severe").

Label shift, worked

A malaria screener

Built in a clinic where 10% of smears are positive. Sensitivity 95%, specificity 95%.

Dry season, 2% positive, 1000 smears: 19 true alarms, 49 false: precision $19/68 = 28\%$.

Rainy season, 30% positive: 285 true, 35 false: precision $285/320 = 89\%$.

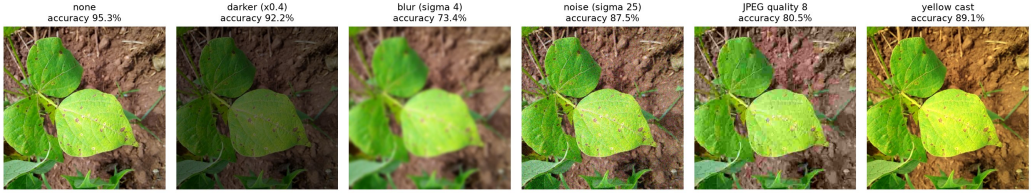
Intuition

The model did not change at all, yet the meaning of a positive result changed threefold. Precision depends on how common the condition is, which depends on place and season.

Practitioner's Note

Report sensitivity and specificity (which do not depend on prevalence), and translate them into precision for the prevalence where the system will be used.

A real test: the bean classifier under corruptions



Week 2's frozen ResNet-50 features + logistic regression on iBean, tested on the same 128 test leaves after six realistic changes. Clean accuracy 95.3%; darker 92.2%; blur 73.4%; sensor noise 87.5%; heavy JPEG 80.5%; yellow cast 89.1%.

Intuition

Nothing about the leaves changed: same plants, same diseases, same labels. Only the *camera* changed, and the model's accuracy moved with it. The size of each drop tells you which changes to test and augment against.

Why do networks break under small changes?

- Networks use *whatever* pattern predicts the label in the training data, including texture, colour and fine detail that humans ignore.
- Blur removes high frequencies (Week 1, Day 3); noise adds them. A model that relies on fine texture is hit by both.
- Features learned on ImageNet carry ImageNet's habits: well-lit, sharp, centred photos.

Intuition

Humans recognise a leaf disease from the *shape and colour* of lesions, robustly. A network may have learned "this exact speckle texture", which disappears with blur. Same accuracy on clean images, very different reasons.

Practitioner's Note

ImageNet-trained CNNs are known to rely heavily on texture rather than shape (Geirhos et al., 2019): a cat-shaped image with elephant-skin texture is called "elephant".

Shortcut learning

- A **shortcut** is a feature that predicts the label in the data but not in the world.
- Documented examples:
  - chest X-ray models detecting the hospital (from text markers and scanner type) rather than the disease;
  - skin-lesion models using the ruler that dermatologists place next to suspicious lesions;
  - "cow" classifiers that fail on cows on a beach, having learned "green pasture".

Intuition

The network is a brilliant, lazy student: it finds the easiest rule that gets full marks. If the easiest rule is the background, the hospital or the ruler, that is what it learns.

Practitioner's Note

Grad-CAM (Week 2, Day 2) and testing on deliberately mismatched images (cows on beaches) are cheap ways to catch shortcuts before deployment.

Finding shortcuts in your own model

1. **Look at attention**: Grad-CAM on 20 correct and 20 wrong predictions. Does the heat sit on the object or the background?
2. **Background swap**: paste objects onto other backgrounds; does the prediction change?
3. **Metadata check**: can a simple classifier predict the label from metadata alone (site, device, date)? If yes, the network may learn it too.
4. **Slice**: evaluate per site, per device, per condition.

A worrying sign

A field-photo classifier reaches 98%. A logistic regression on the *phone model* alone reaches 80%, because each disease was photographed mostly with one phone. The 98% is probably partly a phone detector.

Making models more robust

- **Collect diverse data**: more phones, sites, seasons, people. The best fix, when possible.
- **Augment** with the changes you expect: brightness, blur, noise, compression, colour casts.
- **Larger, more broadly pre-trained models** (CLIP, DINO) are usually more robust.
- **Fine-tune on a little target data**: even 50 labelled images from the new site help a lot.
- **Test-time adaptation**: update normalisation statistics on unlabelled target images.

A robustness table for your capstone

Report clean accuracy and accuracy under 4–6 corruptions relevant to your setting, as on the previous slide. Then retrain with matching augmentation and report the table again. Two tables, one honest story.

Splitting data to measure shift honestly

- **Random split**: train and test share sites, cameras, patients. Measures in-distribution accuracy only.
- **Grouped split**: hold out whole groups: a hospital, a farm, a camera location, a month.
- **Temporal split**: train on the past, test on the future.

A typical gap

Camera-trap classifiers often report well above 90% with random splits but drop by 10–30 points on held-out locations, because backgrounds repeat within a site.

Intuition

The split defines the question. "Will this work on the next photo from the same farm?" and "will this work on a new farm?" are different questions with different answers.

Practitioner's Note

Ask the deployment question first, then design the split to match it.

Part 2

Adversarial examples

Imperceptible changes, wrong answers

- Szegedy et al. (2014): adding a carefully chosen, invisible pattern to an image can make a network output any label with high confidence.
- The pattern is computed with the same tool as training: the gradient, but with respect to the *image*, not the weights.

Fast gradient sign method

$$x_{\text{adv}} = x + \varepsilon \cdot \operatorname{sign}\big(\nabla_x \mathcal{L}(x, y)\big)$$

ε: e.g. 2/255 per pixel, invisible to people.

Intuition

In 150 000 dimensions (a $224 \times 224 \times 3$ image), many tiny nudges, each in the most harmful direction, add up to a large change in the output. A step too small to see in each pixel is a giant step in total.

Adding up tiny changes

A neuron sums $w \cdot x$ over 150 000 pixels. Change every pixel by ± 0.008 in the sign of w; if the average $|w| = 0.01$, the sum moves by $150\,000 \times 0.008 \times 0.01 = 12$.

Physical attacks and real risks

- **Patches and stickers**: printed patterns that make a stop sign read as a speed limit, or make a person invisible to a detector.
- **Typographic attacks**: yesterday's "DOG" label on a cat fooled CLIP with a pen and paper.
- **Glasses, make-up, clothing** designed to fool face recognition.

Practitioner's Note

Defences exist (adversarial training, input purification, certified robustness for small changes) but none is complete. For high-stakes systems, assume a determined attacker can find a failure and design the process so one model error is not catastrophic.

Intuition

Adversarial examples show that networks do not see the way we do: their decisions depend on directions in pixel space that are meaningless to us.

Adversarial training and its trade-offs

- **Adversarial training**: during training, attack each batch and train on the attacked images.
- Gives real robustness to small, bounded perturbations.
- Costs: several times more training compute, and usually a few points lower clean accuracy.
- Robust models' gradients tend to look more like object shapes: they rely less on invisible texture.

Intuition

Training against an attacker removes the most fragile shortcuts. The model gives up a little accuracy to stop relying on patterns humans cannot see.

Practitioner's Note

For most capstones, ordinary corruption robustness (blur, light, compression) matters far more than adversarial robustness. Test those first.

Part 3

Can we trust the confidence?

What calibration means

- A model is **calibrated** if, among all predictions made with 80% confidence, about 80% are correct.
- Confidence = the largest softmax probability.
- Modern deep networks are often **over-confident**: 95% confidence, 85% correct.

Checking ten predictions

Ten predictions each at 90% confidence; seven correct. Accuracy 70% vs. confidence 90%: over-confident by 20 points.

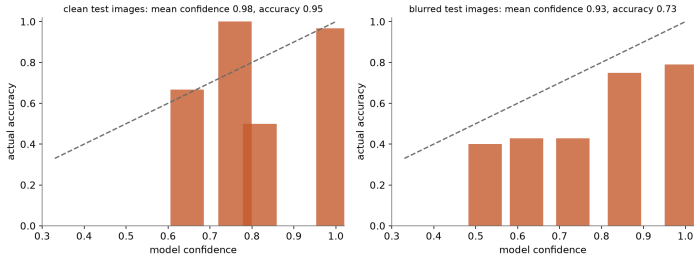
Intuition

A weather forecaster who says "70% chance of rain" is calibrated if it rains on about 70% of those days. We can trust their number, not just their verdict.

Practitioner's Note

Calibration matters whenever a threshold is applied ("refer to a doctor if confidence < 0.9") or probabilities are combined with costs.

Reliability diagrams for the bean classifier



Group predictions by confidence; for each group, plot the actual accuracy (bars). Perfect calibration lies on the dashed diagonal. Clean test leaves: mean confidence 0.98, accuracy 0.95. Blurred leaves: mean confidence 0.93, accuracy 0.73.

Intuition

The key question is whether confidence falls as fast as accuracy when images get worse. If it does not, the model does not know that it has become less reliable, which is the most dangerous kind of error.

Expected calibration error, worked

ECE

$$\mathrm{ECE} = \sum_b \frac{n_b}{N} \left| \mathrm{acc}_b - \mathrm{conf}_b \right|$$

Weighted average gap between accuracy and confidence across confidence bins.

Three bins, 100 predictions

bin	n	conf	acc
low	20	0.55	0.50
mid	30	0.75	0.60
high	50	0.95	0.90

$0.2 \times 0.05 + 0.3 \times 0.15 + 0.5 \times 0.05 = 0.08.$

Intuition

ECE is the average lie, in probability points, that the model tells about its own reliability. 0.08 means: on average, its confidence is off by 8 points.

Practitioner's Note

ECE depends on the number of bins and on test-set size; with only 128 test images, report it with care and always show the diagram.

Fixing calibration: temperature scaling

Temperature scaling

$$p = \mathrm{softmax}(z/T)$$

One number T, fitted on a validation set to minimise cross-entropy. $T > 1$ softens over-confident predictions.

Two classes

Logits $(3, 1)$. With $T = 1$: $p = e^3/(e^3 + e^1) = 0.88$. With $T = 2$: $e^{1.5}/(e^{1.5}+e^{0.5}) = 0.73$. The ranking (and accuracy) is unchanged; only the confidence shrinks.

Intuition

Temperature scaling turns down the volume of the model's certainty without changing what it says. Cheap, effective, and does not touch the weights.

Practitioner's Note

A temperature fitted on clean validation images does not fix over-confidence under shift. Calibrate on data that look like deployment, and re-check after any change of camera or site.

Knowing what you do not know

- **Aleatoric uncertainty**: the image itself is ambiguous (a blurry, half-hidden leaf). More data will not fix it.
- **Epistemic uncertainty**: the model has not seen anything like this (a cassava leaf shown to a bean model). More data would fix it.
- Tools: **ensembles** (train 5 models; disagreement = uncertainty), **Monte Carlo dropout**, **conformal prediction** (output a *set* of labels with a guaranteed coverage).

An ensemble of three

Bean image: models say rust 0.9, 0.85, 0.92: agree, low uncertainty.
Goat image: models say healthy 0.8, rust 0.7, spot 0.9: each is confident, but they disagree: high epistemic uncertainty. Refer it to a human.

Intuition

One confident expert can be confidently wrong; a panel of experts who disagree is a warning sign.

Conformal prediction, worked

1. On a calibration set of n images, compute each image's score $s = 1 - p(\text{true class})$.
2. Take $\hat{q}$ = the $\lceil (n+1)(1-\alpha) \rceil$-th smallest score.
3. For a new image, output every class with $1 - p(\text{class}) \leq \hat{q}$.

$\alpha = 0.1$, 99 calibration images

$\hat{q}$ is the 90th smallest score, say 0.6. A new leaf with $p = (0.55, 0.42, 0.03)$: classes with $p \geq 0.4$: {healthy, rust}. A clear image with $p = (0.95, 0.04, 0.01)$: {healthy}.

Intuition

Instead of one possibly wrong answer, give a short list that contains the right answer 90% of the time, guaranteed for new images like the calibration ones. Bigger sets signal harder images.

Practitioner's Note

The guarantee needs calibration data from the same distribution as deployment. Under shift, recalibrate.

Out-of-distribution detection and the right to abstain

- A bean classifier given a photo of a hand, a goat or a blank wall still outputs one of three bean classes.
- **OOD detection** flags inputs unlike the training data: low maximum probability, high ensemble disagreement, large distance in feature space (e.g. CLIP or DINO features), or high autoencoder reconstruction error (Day 1).
- Then the system can **abstain**: "I cannot assess this image; please retake or consult an expert."

Practitioner's Note

A system that abstains on 10% of images and is right on 97% of the rest is often more useful, and much safer, than one that answers everything and is right on 90%.

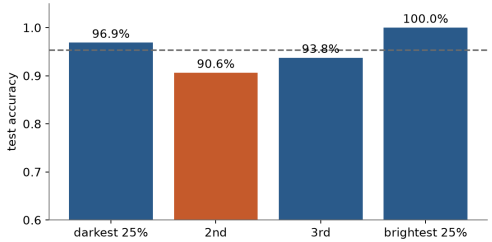
Intuition

Good human experts say "I don't know" and "this needs a specialist". We should build that into our systems deliberately.

Part 4

Fairness: who does the model fail?

An average can hide a weak subgroup



The bean classifier's test accuracy, split by image brightness (quartiles). Overall accuracy: dashed line. Best group 100.0%; worst group 90.6%. Interestingly, the darkest images are *not* the worst.

Intuition

The single headline number is an average over subgroups that may differ a lot. Here the subgroup is lighting; in medicine or face analysis it can be age, sex or skin tone, and the consequences are about people.

Practitioner's Note

With 32 images per group, 90.6% vs. 100.0% is a difference of just 3 images: quite possibly chance. Report group sizes and a confidence interval for each accuracy, e.g. by bootstrapping: resample the test images with replacement many times and recompute.

Gender Shades (2018)

- Buolamwini and Gebru audited three commercial face-analysis systems predicting gender.
- They built a balanced benchmark of parliamentarians' photos from three African and three European countries.
- Error rates: below 1% for lighter-skinned men; up to about 35% for darker-skinned women.
- After publication, the companies improved their systems substantially.

Intuition

The systems were not built to discriminate. Their training and test data contained few darker-skinned women, so nobody measured the problem, and nobody fixed it. What is not measured is not fixed.

Practitioner's Note

Building a careful, balanced evaluation set is a powerful contribution. The Gender Shades benchmark used African faces to expose a problem that affected Africans.

Where unfairness comes from

- **Representation**: some groups are rare in the data (dermatology datasets are dominated by lighter skin).
- **Measurement**: sensors and processing work differently (pulse oximeters and some cameras on darker skin; auto-exposure tuned on lighter faces).
- **Labels**: annotators' errors and assumptions differ by group.
- **Proxy targets**: predicting "cost of care" instead of "need for care" encodes past inequality.
- **Deployment**: who uses the system, on whom, and who can contest its decisions.

Intuition

Bias is not a bug in one line of code; it enters at every stage, from which photos were taken to how results are used. So checks must happen at every stage too.

Practitioner's Note

Week 1, Day 2 showed that a camera is a measuring instrument with its own design choices. Film and early digital colour calibration were tuned on lighter skin; dark skin was poorly exposed for decades.

A case from medical imaging: skin

- Dermatology AI datasets have been dominated by images of lighter skin; studies have found few or no images of darker skin types in many public datasets.
- Conditions can look different on darker skin (redness is less visible, pigmentation changes differ).
- Evaluations on diverse images (e.g. the Diverse Dermatology Images set, 2022) found performance drops for darker skin tones in several published models.

Intuition

The model is only as good as its teachers' examples. If textbooks and datasets show a condition mostly on light skin, both doctors and models learn a narrower picture of the disease.

Practitioner's Note

For any health project: record and report skin type, age and sex distributions of your data, and evaluate on each where possible.

Measuring fairness

- **Disaggregated evaluation**: report every metric per subgroup, with group sizes.
- **Equal error rates**: e.g. similar false-negative rates across groups for a screening test.
- **Worst-group accuracy**: optimise and report the weakest group, not just the average.

A screening test

Group A: 200 sick, 190 detected (FNR 5%). Group B: 50 sick, 40 detected (FNR 20%). Overall: $230/250 = 92\%$ sensitivity. Group B's patients are missed four times as often.

Intuition

Different fairness definitions can contradict each other mathematically; you cannot satisfy them all at once. Choosing which one matters is a decision about values, made with the people affected, not a purely technical choice.

Practitioner's Note

At minimum: report per-group results, state which groups you could *not* evaluate, and say so in your model card.

Mitigating unfairness

- **Data**: collect more from under-represented groups, with consent; this usually helps most.
- **Training**: re-weight or over-sample weak groups; optimise worst-group loss.
- **Post-processing**: different thresholds per group where justified and lawful.
- **Process**: involve affected communities in design and evaluation; allow people to contest decisions.
- **Do not build it**: some applications should not exist.

Practitioner's Note

Removing a sensitive attribute from the input rarely removes bias: skin tone, clothing, background and image quality all act as proxies. Measure outcomes instead of assuming fairness through blindness.

Global data gaps



Bdx, CC BY-SA 4.0

- Large vision datasets are dominated by images from North America and Europe; studies of ImageNet and Open Images found a small share of images from Africa.
- Models recognise “a wedding”, “a kitchen” or “soap” better in the settings they saw most.
- Local objects (shared taxis, moto-taxis, local crops, markets, dress) are rare or mislabelled.

Intuition

A model trained on the world’s photos is really trained on the photos of the people who post most. Datasets made in Cameroon, by people who know the context, change what models can see.

Part 5

Privacy, surveillance, consent and deepfakes

Images are personal data

- Faces, number plates, homes, medical scans, handwriting and even gait can identify people.
- Metadata (EXIF: GPS location, time, device) travels with photos unless removed.
- Combining images with other data (social media, records) reveals far more than either alone.

Basic practice for your projects

- Collect only what you need; explain why to people photographed.
- Strip metadata; blur faces and plates when not needed.
- Store data securely; limit access; delete on schedule.
- Follow the dataset's licence and the law where data were collected.

Intuition

Detection and segmentation (Week 2) are also privacy tools: the same detector that finds faces can blur them automatically.

Face recognition and surveillance

- Face recognition matches a face against a database: phone unlock (1:1, with consent) is very different from scanning crowds against a watch list (1:many, without consent).
- Documented harms: wrongful arrests after false matches, tracking of protesters and journalists, chilling of free assembly.
- Several cities and the EU's AI Act restrict real-time remote biometric identification in public spaces.

The base-rate problem

Scan 100 000 faces a day against a list of 100 wanted people; 10 of them pass by. A system with 99% sensitivity and a 0.1% false-match rate catches about 10, and falsely flags about 100 innocent people.

Intuition

Day 1's anomaly arithmetic, now with people: rare targets plus mass scanning means most alarms are false, and each false alarm is a person stopped.

Consent and community data

- "Publicly visible" is not the same as "consented to being in a dataset". Several face datasets scraped from the web were withdrawn after criticism.
- Communities, not only individuals, have interests in data about them: farming practices, sacred sites, health of a village.
- **Data sovereignty**: who controls data collected in Cameroon, where is it stored, and who benefits from models built on it?

Practitioner's Note

Before collecting images for a project: obtain ethical approval where required (any human subjects, patients, minors); obtain informed consent in a language people understand; agree with partners on data ownership, sharing and credit.

Intuition

Respect is part of good science: the farmers, patients and communities who provide images should share the benefit and the credit.

Deepfakes and synthetic media

- Day 2's generators and Day 1's autoencoders make it easy to swap faces, fabricate photos and clone voices.
- Most deepfake harm documented so far is non-consensual intimate imagery, overwhelmingly targeting women; also fraud and election disinformation.
- "Liar's dividend": genuine evidence can be dismissed as fake.

What helps

- Provenance and content credentials attached at capture.
- Watermarking of generated content by major tools.
- Fast takedown processes and laws against non-consensual imagery.
- Media literacy: ask where an image came from.

Practitioner's Note

Deepfake detectors are an arms race; treat their outputs as one piece of evidence, never as proof.

Dual use

- Almost every technique in this course can help or harm:
  - tracking cars to plan roads, or to follow a person;
  - counting crowds for safety, or to identify protesters;
  - recognising animals for conservation, or for poaching;
  - generating images for teaching, or for fraud.

Intuition

The question is rarely "is this technique ethical?" but "is this *application*, by these people, with these safeguards, ethical?".

Practitioner's Note

Questions to ask of any project: Who benefits? Who could be harmed? Who decides? Can those affected contest the outcome? What happens if it is misused?

Part 6

Responsible practice

Documentation: model cards and datasheets

Model card (Mitchell et al., 2019)

- Intended use and out-of-scope uses.
- Training data; evaluation data and how they differ.
- Metrics overall *and per subgroup*; robustness tests.
- Known limitations and ethical considerations.

Datasheet for a dataset (Gebru et al., 2021)

- Who collected it, why, how, with what consent.
- What is in it and what is missing; known biases.
- Licence; recommended and discouraged uses.

Practitioner's Note

Add a short model card (half a page) as an appendix to your one-page capstone summary. It is short, it forces honest thinking, and reviewers and employers recognise it.

Intuition

A model card is a medicine leaflet for a model: what it is for, how well it works, for whom, and what side effects to watch for.

A model card for the bean classifier

Intended use	Teaching example; research on bean leaf disease in Uganda. Not for advice to farmers without field validation.
Model	Frozen ImageNet ResNet-50 features + logistic regression.
Data	iBean (Makerere AI Lab), phone photos from Ugandan fields; 3 classes.
Results	Test accuracy 95.3% (128 images). By brightness quartile: 90.6% to 100.0%.
Robustness	Darker 92.2%, blur 73.4%, noise 87.5%, JPEG 80.5%, yellow cast 89.1%.
Calibration	Clean: confidence 0.98, accuracy 0.95; blurred: confidence 0.93, accuracy 0.73.
Limitations	Three classes only; other diseases, deficiencies and crops will be forced into one of them. Not tested in Cameroon, other seasons or other phones.
Ethics	A wrong "healthy" could delay treatment; keep an extension officer in the loop.

Intuition

Every line of this card came from an experiment run in this course. That is what a model card should be: evidence, not marketing.

Regulation and governance

- **EU AI Act** (2024): risk-based; bans some uses (e.g. untargeted scraping of facial images, emotion recognition at work and school); strict duties for high-risk systems (medical, hiring, policing, critical infrastructure). Applies to systems used in the EU, wherever they are built.
- **African Union Continental AI Strategy** (2024): development-focused, emphasises skills, data, infrastructure and ethics.
- **National data-protection laws**: many African countries, including Cameroon, now have or are adopting laws on personal data; the Malabo Convention covers data protection at AU level.
- **Sector rules**: medical devices, aviation, vehicles.

Practitioner's Note

Laws change quickly. For any real deployment, check the current law of the country where data are collected and where the system is used, and involve a legal or ethics adviser.

Intuition

Regulation tends to follow the same logic as this lecture: higher stakes $\rightarrow$ more evidence of accuracy, robustness, fairness and human oversight.

Humans in the loop

- **Triage**: the model sorts cases by urgency; humans decide (chest X-rays, camera traps).
- **Second reader**: the model checks the human's work.
- **Abstention**: uncertain cases go to humans automatically.

Practitioner's Note

Automation bias: people over-trust machine outputs, even when the machine is wrong. Show uncertainty, show the evidence (heatmaps, boxes), and train users on failure cases.

Intuition

"A human checks every output" is only a safeguard if the human has the time, information and authority to disagree. Otherwise it is a signature on the machine's decision.

Explanations: useful, but limited

- Heatmaps (Grad-CAM, occlusion; Week 2) show *where* the model looked, not *why* or *what* it used there.
- Different explanation methods can disagree on the same prediction.
- Some heatmap methods look similar even for a randomly initialised network (Adebayo et al., 2018): always sanity-check.

Intuition

An explanation is a hypothesis about the model, not a confession from it. Test it: if the model really uses the lesion, covering the lesion should change the prediction.

Practitioner's Note

Use explanations to *debug* (find shortcuts), and be cautious about using them to *justify* decisions to patients or farmers.

Monitoring after deployment

- **Input drift**: track simple image statistics (brightness, blur, size, device) and feature-space distances from training data.
- **Output drift**: class frequencies and confidence distributions over time.
- **Ground truth**: label a small random sample every month; measure real accuracy.
- **Feedback channel**: users can report errors easily.

An early warning

A disease app's share of "healthy" predictions jumps from 40% to 75% in a week. Either the season changed, or a phone update changed the camera processing. Either way: investigate before trusting the numbers.

Environmental and economic costs

- Training and serving large models consume electricity, water for cooling, and hardware with mined materials.
- Costs and benefits are unevenly distributed: compute is concentrated in a few countries and companies.
- Efficient models (YOLO-n, MobileNet, linear probes on frozen features) often deliver most of the value at a tiny fraction of the cost.

This course

Every model trained in this course ran on a 2-core laptop CPU. The bean probe trains in seconds and reached 95.3%; fine-tuning a ResNet reached 95% in a few minutes.

Intuition

Frugal AI is not second-best AI. Choosing the smallest model that meets the need is good engineering, good economics, and good for the planet.

A responsible-deployment checklist

1. Is the problem worth solving with vision, and do affected people want it?
2. Is the data collected lawfully, with consent, and representative of users?
3. Is the test split grouped to match deployment?
4. Are results reported per subgroup, with uncertainty?
5. Is robustness tested against expected changes?
6. Is confidence calibrated, with a way to abstain?
7. Is there a human with time and authority to override?
8. Is there a model card and a datasheet?
9. Is there monitoring and a feedback channel?
10. What is the plan if it is misused or fails?

Practitioner's Note

Use this list for your capstone's limitations section: answer each item honestly in one or two sentences, including "not done, and here is why".

Case study: designing a malaria smear assistant responsibly

Question	Design choice
Who uses it?	laboratory technicians in district hospitals; final decision stays with them
Data	smears from several regions, stains, microscopes and seasons; patient consent and ethics approval
Split	by hospital; held-out hospitals for the final test
Metric	sensitivity first; per-hospital and per-stain results; precision at local prevalence
Uncertainty	poor-quality smears flagged as "unreadable, prepare a new slide"
Interface	boxes on each detected parasite so the technician can verify quickly
Monitoring	monthly random audit by an expert microscopist
Failure plan	system can be switched off; manual reading remains the standard

Intuition

Every row uses something from this course. Responsible design is not an extra module; it is the course applied with care.

Part 7

The road ahead

What we covered in three weeks

Week 1: foundations
human vision; images as data; filtering and frequency; edges, corners, features; from engineered to learned features

Week 2: deep vision
CNNs and training; architectures and transfer; detection; segmentation; transformers, captioning, open vocabulary

Week 3: creating and connecting
autoencoders and VAEs; diffusion; CLIP and vision–language; applications; robustness, fairness and ethics

Intuition

The same few ideas kept returning: images are measurements; useful features are invariant; convolution and attention share information across space; learning replaces hand design; and evaluation must match the real question.

Open problems in computer vision

- **Robustness**: models that hold up across cameras, places and time.
- **Data efficiency**: learning from a few examples, as children do.
- **3D and physical understanding**: shape, depth, physics, cause and effect.
- **Video and time**: long videos, actions, intentions.
- **Reasoning**: counting, relations, "why" questions.
- **Trustworthy uncertainty**: knowing when not to answer.
- **Efficiency**: strong models on phones and microcontrollers.
- **Fair and representative data**: for the whole world, in its languages.

Intuition

Many of these are where today's large models still fail, and several can be studied with modest compute: careful evaluation, new datasets and efficient methods do not need a data centre.

Trends to watch

- **Foundation models** for vision (DINO, SAM, CLIP families) as the default starting point, with domain-specific versions for medicine, satellites and agriculture.
- **Multimodal models** that read, see, listen and act.
- **World models**: video generators that learn how scenes evolve, used for planning in robotics.
- **3D**: neural radiance fields and Gaussian splatting turn photos into explorable 3D scenes.
- **Embodied AI**: robots that learn from video and language.
- **Edge AI**: capable models on cheap devices, offline: important where connectivity is limited.

Practitioner's Note

Trends move fast; fundamentals do not. Convolution, attention, losses, evaluation and data hygiene from this course will still matter when today's model names are forgotten.

Computer vision in Africa



Photokadaffi, CC BY-SA 4.0

- Active communities: Deep Learning Indaba, Data Science Africa, Masakhane (language), Black in AI, and AIMS itself.
- African datasets that changed what models see: iBean and iCassava (Makerere), malaria microscopy, African satellite land cover, local language OCR efforts.
- Problems with local priority and global relevance: crop health, maternal health, road safety, climate, conservation, accessibility.

Intuition

The most valuable contribution is often not a bigger model but the right question, the right data, and an honest evaluation in a context that the rest of the field has ignored.

How to keep learning

- **Build**: turn your capstone into a small paper, a dataset, or an open-source tool.
- **Read**: one paper a week; start with the abstract, figures and conclusion; re-implement one figure.
- **Courses**: Stanford CS231n, MIT 6.8300, the free fast.ai course; Szeliski's *Computer Vision: Algorithms and Applications* (free online).
- **Communities**: Indaba, workshops at CVPR/ICCV/NeurIPS (e.g. CV4Africa, Computer Vision for Global Challenges), Kaggle and Zindi competitions.

Practitioner's Note

Keep a "lab notebook": for every experiment, what you tried, what you expected, what happened. It is the habit that separates people who learn quickly from people who repeat mistakes.

Intuition

You now know enough to read most vision papers and to build useful systems. The rest is practice, curiosity and good questions.

Part 8

Wrapping up

Key takeaways

1. Deployment shifts the data; test under realistic corruptions and with grouped splits (our beans: 95.3% clean, 73.4% blurred).
2. Networks take shortcuts and can be fooled by invisible or physical perturbations.
3. Check calibration; add abstention and OOD detection.
4. Averages hide weak subgroups; report per-group results with sizes.
5. Images are personal data; consent, privacy and dual use must shape what we build.
6. Document with model cards and datasheets; monitor after deployment; keep empowered humans in the loop.
7. Frugal, well-evaluated models and good local data are a powerful way to contribute.

Before presentation day

Capstone: final steps

- Robustness table: clean + 4–6 relevant corruptions.
- Per-group or per-condition results where possible.
- Reliability diagram if your model outputs confidences.
- A short model card appended to the one-page summary.
- Rehearse the 10-minute talk at least twice, timed.
- Code frozen this evening.

Tomorrow

- Capstone presentations: 10 minutes + 5 minutes of questions per team
- Quiz 3 in the tutorial slot
- Course feedback

Intuition

Lead your talk with the problem and the people it affects, then the evidence. The best talks are honest about what did not work.

Credits

iBean dataset: Makerere AI Lab and NaCRRI, Uganda (MIT licence). Robustness, calibration and subgroup experiments were run for this lecture on a laptop CPU with `Week3/tools/make_week3_figures.py`. Gender Shades figures from Buolamwini and Gebru (2018), *Proceedings of Machine Learning Research* 81.

Photographs from Wikimedia Commons:
`yaounde_taxis`: Bdx, CC BY-SA 4.0
`moto_mbouda`: Photokadaffi, CC BY-SA 4.0