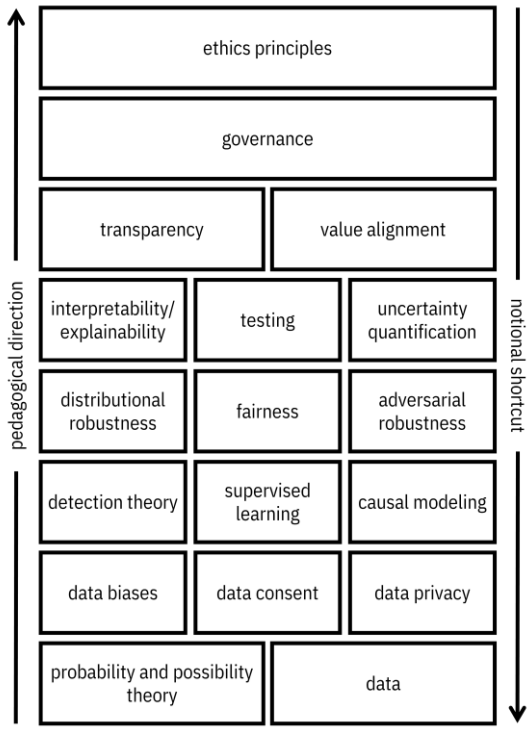
Shortcut

Even though I have admonished you throughout the entire book to slow down, think, and not take shortcuts, I know some of you will still want to take shortcuts. Don't do it. But if you're adamant about it and are going take a shortcut anyway, I might as well equip you properly.

Here is a picture showing how I structured the book, going from bottom to top. This direction makes sense pedagogically because you need to understand the concepts at the bottom before you can understand the nuances of the concepts that are higher up. For example, it is difficult to understand fairness metrics without first covering detection theory, and it is difficult to understand value elicitation about fairness metrics without first covering fairness. However, if you want to jump right into things, you should notionally start at the top and learn things from below as you go along.



Accessible caption. A stack of items in 8 layers. Top layer: ethics principles; layer 2: governance; layer 3: transparency, value alignment; layer 4: interpretability/explainability, testing, uncertainty quantification; layer 5: distributional robustness, fairness, adversarial robustness; layer 6: detection theory, supervised learning, causal modeling; layer 7: data biases, data consent, data privacy; bottom layer: probability and possibility theory, data. An upward arrow is labeled pedagogical direction. A downward arrow is labeled notional shortcut.

The ultimate shortcut is to give you a recipe to follow.

Preparation Steps:

1. Assemble socioculturally diverse team of problem owners, data engineers and model validators including members with lived experience of marginalization.
2. Determine ethics principles, making sure to center the most vulnerable people.
3. Set up data science development and deployment environment that includes fact flow tool to automatically collect and version-control digital artifacts.
4. Install software libraries in environment for testing and mitigating issues related to fairness and robustness, and computing explanations and uncertainties.

Lifecycle Steps:

1. Identify problem.
2. Conduct facilitated participatory design session including panel of diverse stakeholders to answer the following four questions according to ethics principles:
    a. Should the team work on the problem?
    b. Which pillars of trustworthiness are of concern?
    c. What are appropriate metrics?
    d. What are acceptable ranges of metric values?
3. Set up quantitative facts for the identified pillars of trustworthiness and their metrics.
4. If the problem should be worked on, identify relevant dataset.
5. Ensure that dataset has been obtained with consent and does not violate privacy standards.
6. Understand semantics of dataset in detail, including potential unwanted biases.
7. Prepare data and conduct exploratory data analysis with a particular focus on unwanted biases.
8. Train machine learning model.
9. Evaluate model for metrics of trustworthiness of concern, including tests that cover edge cases. Compute explanations or uncertainties if of concern.
10. If metric values are outside acceptable ranges, try other data, try other learning algorithms, or apply mitigation algorithms until metric values are within acceptable ranges.
11. Deploy model, compute explanations or uncertainties along with predictions if of concern, and keep monitoring model for metrics of trustworthiness of concern.