Toward Responsible AI Planning and Decision Making

CS4246/CS5446

AI Planning and Decision Making

Sem 1, AY2021-22

Topics

- Human judgment and irrationality (16.3.4)
  - Cognitive biases and psychology traps
  - How to make better decisions?
- Toward human-aware AI
  - Judgmental heuristics and biases
  - Social, economic, environmental trends
- Toward responsible AI
  - Privacy, Fairness, Transparency
  - Ethical, governance, and regulatory considerations

Types of Decision Theory

- Normative decision theory
  - Describes how ideal, rational agents should behave
    - Explains how a rational agent should make decisions
    - Proposes models of preference and rational decision making
    - Preference ordering, maximum expected utility
- Descriptive decision theory
  - Describes how actual agents (humans) really behave
    - Explains how we make decisions
    - Identify potential pitfalls and biases
    - Prospect Theory, Naturalistic decision making
- Prescriptive decision theory
  - Prescribes guidelines for agents to behave rationally
    - Explains how we should make rational decisions
    - Provides guidelines and tools to support decision making
    - Decision analysis and modeling

Rational Decision Making Revisited

- Are we rational?
  - Why is rational decision making difficult?
  - Will rational decisions and approaches always lead to good outcomes?
  - Are we usually rational?
- Humans are “predictively irrational!”
  - Biases emanating from cognitive heuristics
  - Representativeness
  - Availability
  - Anchoring and adjustments

Garfield the Cat

- Consider:
  - You have a beautiful cat named Garfield.
  - You love your cat; he is your best friend and companion.
  - Somebody offers you $100,000 to buy Garfield.
  - With the money, you could realize many of your dreams, including going overseas for further studies.
  - But you know that Garfield will not survive away from you, and you can't bear to lose him.
- What will you do?

Bounded Rationality

- Views individuals as attempting to make rational decisions
- Acknowledges problem solver:
  - often lack important information on problem situation
  - face time and cost constraints that limit quality and quantity of available information;
  - have limited memory capacity; and
  - are limited by intelligence and perceptions to accurately calculate optimal choice
- In real life, problem solvers satisfice rather than optimize
  - Forego the best solution in favor of one that is acceptable or reasonable
  - Do not examine all alternatives
  - Search for a solution that meets an acceptable level of performance

Judgmental Heuristics

Power of being human …

Judgment

- What is judgment?
  - Major class of cognitive activity, sf. learning and perception
  - An opinion about what is or will be the status of some aspects of the world
- Types of judgment:
  - Prediction
  - Evaluation
  - Determination
  - Likelihood
- Accuracy:
  - Depends on extent to which the mind mirrors environment it attempts to predict

Judgmental Heuristics

- What are heuristics?
  - Simplifying strategies or rules of thumb
  - Simple ways of dealing with complex world
  - Usually provide acceptable results
- Pitfalls:
  - Adopt heuristics without being aware of them
  - Misapplication to inappropriate situations leading to wrong/bad solutions

The Availability Heuristic

- Assess likelihood of event by degree to which instances of event are readily “available” in memory
- Availability of an event in memory depends on:
  - emotional vs. unemotional
  - vivid vs. bland
  - easily imagined vs. difficult to imagine
  - specific vs. vague
- Questions:
  - Why is it a good heuristic?
  - Why is it fallible?

Availability Heuristic in Practice

- Examples:
  - Your Assignment 2 is due in 10 days time. Do you think you will actually finish the assignment by the target date?
  - When was the last time you talked to your boyfriend/girlfriend?
  - When was the last time you went to the post office?
  - In an arbitrary English word, is it more likely to find the letter R in the first position or in the third position?

The Representativeness Heuristic

- Assess event likelihood by similarity to stereotypes
- Examples:
  - How do you classify a new plant?
  - My friend Charlie is outgoing, outspoken, adventurous, and hopes to become an Air-Force pilot. Do you think Charlie is a man or a woman?
  - What is the likely diagnosis of a patient with fever, running nose, cough and sore throat?
- Questions:
  - Why is it a good heuristic?
  - Why is it fallible?

Anchoring and Adjustment

- Starting from initial value, adjusting to yield final decision
- Examples:
  - If you had score full marks in Test 1 of a difficult module, would you expect to do as well in Test 2 and Test 3 of the same module?
  - If you had a good meal at a new restaurant, would you go back again?
  - “Love at first sight”
- Why is it a good heuristic?
- Why is it fallible?

Cognitive Biases and Psychological Traps

Danger zones in decision making and how to avoid them

Hidden Traps in Decision Making

- Recallability trap
- Anchoring trap
- Confirming evidence trap
- Framing trap
- Overconfidence trap and Prudence trap
- Insensitive to sample size trap
- Status-quo trap
- …

Examples:

- Which of the following causes more death in Singapore in 2019?
  - Pneumonia
  - Accidents

- Did you address the Lecturer Sir/Madam in the first email message you send to your new CS4246/5446 Lecturer?

1. Recallability trap

- What is the trap?
  - Giving undue weight to recent, dramatic events
  - Based upon vividness and recency
  - An event whose instances are more easily recalled will appear more numerous than an event of equal frequency whose instances are less easily recalled
- What can you do about it?
  - Carefully examine assumptions to ensure they are not overly influenced by recent memory
  - Get actual statistics if possible
  - Do not be guided by impressions

Examples

- A newly hired programmer for a software house has four years of experience and good all-around qualifications. When asked to estimate the starting salary for this employee, my friend (knowing very little about the profession or the industry) guessed an annual salary of $48,000. What is your estimate?

- You ordered food delivery from a famous restaurant. But you suffered mild food poisoning after dinner that night. Would you order from or visit the same restaurant again?

9. Anchoring trap

- What is the trap?
  - Giving disproportionate weight to the initial information received
  - Adjustments away from anchors usually not sufficient to change effects of anchor
  - Answers are biased toward the initial anchor, even if it is irrelevant
  - Different starting points yield different answers

- What can you do about it?
  - View problem from different perspectives
  - Be open-minded
  - Think of problem first before consulting others
  - Be careful to avoid anchoring others' opinions
  - Be wary of anchors in discussions and negotiations

Example: Gains vs. Losses

- An unfortunate pandemic hits a poor country, and due to resource constraints, three emergency hospital camps need to be closed and leaving 6000 patients without care. Consider the two alternatives:
- Plan A:
  - This plan will save one of the three camps and continue to care for 2000 patients
- Plan B:
  - This plan has a 1/3 probability of saving all three camps and providing care to all 6000 patients, but has a 2/3 probability of saving no camps and providing no care for any patients
- Which plan will you choose?

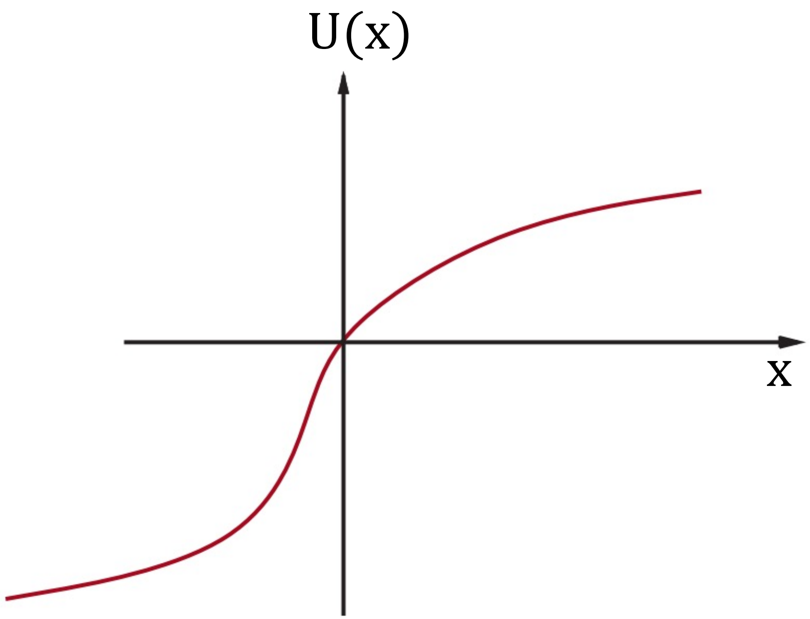
Example: Gains vs Losses (Cont.)

- Consider the following alternatives instead:
- Plan C:
  - This plan will result in the loss of two of the three camps and cutting care for 4000 patients
- Plan D:
  - This plan has a 2/3 probability of resulting in the loss of all three camps and leaving all 6000 patients without care, but has a 1/3 probability of losing no camps and leaving no patients without care
- Now which plan will you choose?

Empirical Results and Implications

- Note that:
  - The two sets of alternatives are objectively …
  - Changing outcome descriptions from patients and camps saved to those lost sufficient to shift prototypical choice from risk-averse to risk-seeking behavior
  - Individuals treat risks concerning perceived gains differently from risks concerning perceived losses
  - Outcomes evaluated relative to a neutral reference point which may be different in different frames
  - Prospect Theory – Kahneman and Tversky, 1979 (Nobel prize work)

Prospect Theory



Source: RN Figure 16.2b

- Reference- or framing-dependent utility functions
- Changes, not levels, in measures of "goodness" matter
- Gains are different from losses
  - Loss aversion
- Risk preferences
  - Risk aversion over gains
  - Risk seeking over losses
- Diminishing sensitivity
- Probability weighting

Kahneman, D. and A. Tversky, *Prospect Theory: An Analysis of Decision under Risk.* Econometrica, 1979. **47**(2): p. 263-291.

Framing trap

- What is the trap?
  - Occurs when a problem is misstated, undermining entire decision making process
  - Presentation of information can significantly impact decision making under uncertainty
  - Intuitions about risk or uncertainty routinely deviate from rationality because of misunderstanding of
    - nature of uncertainty; and
    - effects of framing
  - Often closely related to other traps; a frame can:
    - establish the status-quo or introduce an anchor
    - highlight sunk cost or lead toward confirming evidence

Framing Trap

- What can you do about it?
  - Do not automatically accept the initial frame; always reframe problem in various ways; look for distortions
  - Pose problems in a neutral, redundant way that combines gains and losses and uses different reference points
  - Consider framing effects throughout decision process
  - Examining others' decision framing; challenge with different frames

5. Insensitive to sample size trap

- What is the trap?
  - While sample size is a fundamental concept in statistics, it is rarely a part of our intuitions
  - People tend to use the representativeness heuristic when responding to problems dealing with sampling
- What can you do about it?
  - Carefully recall the sample size on which judgment is based
  - Avoid generalization based on small sample size or limited personal experience
  - Find the real statistics if possible; at least discuss with colleagues with similar experiences to calibrate judgment

Status-quo trap

- What is the trap?
  - Maintaining current situation despite better alternatives
  - “Let’s not rock the boat right now” mentality
  - To protect egos and preference for less psychological risk
  - Breaking status-quo means taking action, and taking responsibility, thus opening up to criticism and regret
  - The more choices given, the more pull the status quo has
    - One alternate treatment versus two or more
- What can you do about it?
  - Status-quo may be the best choice, but be careful about choosing it for the wrong reason
  - Examine how status-quo would serve the main objectives
  - Identify barriers to change
  - Identify other options
  - Avoid over-emphasizing efforts or costs in changing
  - Evaluate desirability with respect to time
  - Force a decision on alternatives if appropriate

11. Overconfidence and Prudence Traps

- What is the trap?
  - Overestimating accuracy of our forecasts
    - May lead to the sunken cost trap - refuse to withdraw from a losing situation, or to continue to put in money, effort, time and other resources after bad investments.
  - Overcautious when estimating uncertain events

- What can you do about it?
  - Make forecasts and judge probabilities systematically
  - Consider the limits of possible range of values
  - Avoid being anchored by initial estimates
  - Imaging circumstances outside the possible range
  - Challenge the estimates
  - Challenge estimates of others
  - State estimates honestly and explain limitations
  - Test estimates over a range to assess impact
  - Re-examine the more sensitive estimates

12. Confirming evidence trap

- What is the trap?
  - Seeking out information supporting an existing prediction and to discount opposing information
  - Seek confirmatory evidence and do not search for disconfirming information for decision processes
  - Search for challenging, or disconfirming, evidence will actually provide the most useful insights
- What can you do about it?
  - Always examine all evidence with equal rigor
  - Get devil's advocate and build counterarguments
  - Be honest about the motives
  - When seeking advice, do not ask leading questions that would lead to confirming evidence

Summary

- Using judgmental heuristics:
  - Loss in quality of solutions usually outweighed by time saved
- Dangers of using heuristics:
  - Situations in which the loss in quality of solutions outweighs time saved
  - Net aware of using heuristics and do not know if they are appropriate
- Other types of biases:
  - Under risk
  - In sequential decisions
  - As hindrance to creativity

In general …

- Judgmental or behavioral decision making
  - Important , emerging field, complementing evidence based and classic utility-based decision making
  - Understanding human behavior from systematic studies and experiments and theoretical frameworks
  - Applications in prediction, diagnosis, intervention, management, policy making, and public communication (sf. Behavioral Economics)
  - Directly relevant to the study of ethical, governance and regulatory considerations in AI
- To improve decision making capabilities (of human-aware AI):
  - Identify common judgmental heuristics
  - Understand potential adverse effects and biases
  - Selectively and correctly apply heuristics and insights

"It is concluded that AI has not yet been impactful against COVID-19. Its use is hampered by a lack of data, and by too much data. Overcoming these constraints will require a careful balance between data privacy and public health, and rigorous human-AI interaction."

Artificial Intelligence against COVID-19: An Early Review - **Wim Naudé,** IZA Institute of Labor Economics, 2020

Toward Human-Aware AI

- Main approaches:
  - "Robust and beneficial" AI with clear social benefits [1] that can serve as cognitive orthoses or prostheses [2] for the decision makers and actors.
  - Disruptive AI that improves, leverages, and extends human cognition and capability to make better decisions leading to better outcomes in dynamic situations.

[1] FORD, K.M., HAYES, P.J., GLYMOUR, C., and ALLEN, J., 2015. Cognitive Orthoses: Toward Human-Centered AI. In AI Magazine, 5-8. DOI= http://dx.doi.org/http://dx.doi.org/10.1609/aimag.v36i4.2629.

[2] RUSSELL, S., DIETTERICH, T., HORVITZ, E., SELMAN, B., ROSSI, F., HASSABIS, D., LEGG, S., SULEYMAN, M., GEORGE, D., and PHOENIX, S., 2015. Research Priorities for Robust and Beneficial Artificial Intelligence: An Open Letter. In AI Magazine, 3. DOI= http://dx.doi.org/http://dx.doi.org/10.1609/aimag.v36i4.2621

Social, Economic, and Environmental Trends

- A changing world:
  - COVID-19 pandemic response and recovery, climate change, sustainability, equality, diversity, inclusiveness …
  - Climate changes
  - United Nations Development Goals
- Industrial Revolution 4.0
  - Digitization integrated with physical and biological systems to enable contextualized, customized and personalized production of goods and services
- New technological and innovations
  - Internet of things, big data, deep learning, quantum computing

Impact on AI Planning and Decision Making

- Desiderata of human aware AI
  - Contextualized, customized, personalized
  - Integrate cognitive psychology and neuroscience findings
  - Detect and manage changes
- Desiderata of responsible AI
  - Consider social, technical, economic, and environmental conditions
  - Consider ethics, governance, and regulatory issues

Responsible AI

Privacy, fairness, transparency
Ethical, governance and regulatory considerations

Toward Responsible AI

- Incorporating ethical, governance, and regulatory considerations into AI systems design, engineering and use
  - What is responsible AI? Why does it matter?
  - What are the main characteristics of responsible AI?
  - How to develop and apply responsible AI?
- Focus: Human-Aware AI for Good
  - Beneficial AI working for, working with, working alongside humans

Common Principles

How to incorporate these considerations into "rational" MEU decision making?

- Ensure safety
- Respect privacy
- Ensure fairness
- Promote trust
- Establish accountability
- Provide transparency
- Attribute responsibility
- Reflect diversity/inclusion
- Support equality
- Facilitate collaboration
- Uphold human rights and values
- Limit harmful uses of AI
- …

Ref: AIMA4e, Chapter 27

Example: Privacy and Re-identification

Name	Gender	Age	Postal Code	Smoker	Diagnosis
Mei Mei	F	68	81317	N	Dementia
Susan	F	61	81382	N	Gastric Disease
Lee	F	67	81304	N	Arthritis
Seng	M	53	81359	Y	Heart Disease
James	M	59	81303	Y	Kidney Disease
Lily	F	52	81359	N	COVID-19
Tony	M	59	81330	Y	Diabetes, Hypertension
Cindy	F	55	81344	N	Lung Cancer

Source: Adapted from The Ethical Algorithm, Chapter 1, 2019

Example: Will K-Anonymity Help?

Suppressed Coarsened Coarsened

Name	Gender	Age	Postal Code	Smoker	Diagnosis
*	F	60-70	813**	N	Dementia
*	F	60-70	813**	N	Gastric Disease
*	M	60-70	813**	N	Arthritis
*	M	50-60	813**	Y	Heart Disease
*	M	50-60	813**	Y	Kidney Disease
*	F	50-60	813**	N	COVID-19
*	M	50-60	813**	Y	Diabetes, Hypertension
*	F	50-60	813**	N	Lung Cancer

Source: Adapted from The Ethical Algorithm, Chapter 1, 2019

Problems with K-Anonymity and Others

- Re-identification is not the only privacy risk.
- Suppose we know that Lily is a 50+ patient at hospital A – either COVID-19 or lung cancer
- Still a serious privacy violation, cannot be prevented by k-anonymity
- Guarantees go away when multiple datasets are released – Lily went to both hospital A and B

Name	Gender	Age	Postal Code	Smoker	Diagnosis (A)
*	F	50-60	813**	N	COVID-19
*	F	50-60	813**	N	Lung Cancer

Name	Gender	Age	Postal Code	Smoker	Diagnosis (B)
*	F	50-60	813**	N	COVID-19
*	F	50-60	813**	N	Pancreatic Cancer
*	F	50-60	813**	N	Allergies

Source: Adapted from The Ethical Algorithm, Chapter 1, 2019

Privacy

- Assumption:
  - Cybersecurity and secured systems protocols in place
- Data collection protocols:
  - HIPAA - Health Insurance Portability and Accountability Act of 1996 (USA)
  - FERPA – Family Educational Rights and Privacy Act of 1974 (USA)
  - GDPR - The General Data Protection Regulation 2016/679 (EU)
  - PDPC - The Personal Data Protection Act 2012 (Singapore)
- Methods (sharing de-identified data in central database)
  - De-identification
  - Generalizing fields
  - K-anonymity
  - Aggregate querying
  - Limiting multiple queries
  - Differential privacy (stronger guarantee)
- Methods (protected sharing with or without central database)
  - Federated learning

Ref: AIMA4e, Chapter 27

Making Privacy Preserving Decisions

- Main issue:
  - Value gained from sharing data balanced against individual’s right to privacy
- Examples:
  - population censor data, handphone usage data, medical research data
- If you are a developer, user, or owner/manager/regulator:
  1. What is privacy?
  2. What can be done to ensure privacy?
  3. What is the trade-off between accuracy and privacy?
  4. What are the implications?
  5. Who should make the decisions? When?

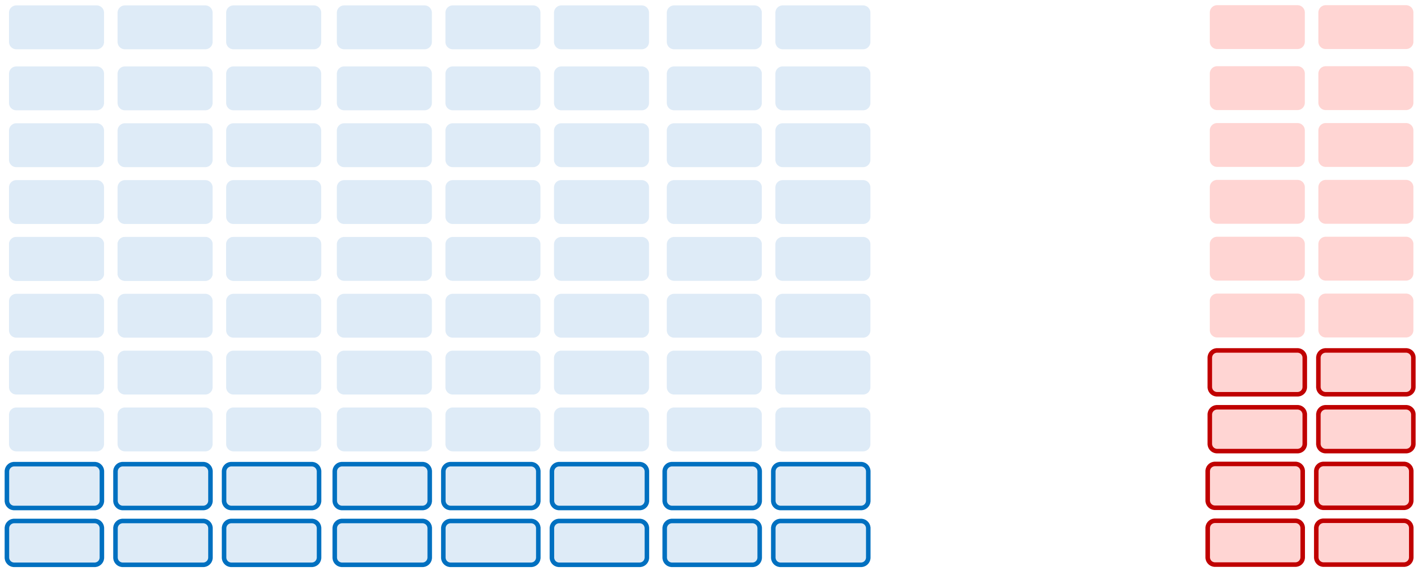
Example: Fairness in Context

- Consider 100 loan applications. 80 applicants have blue nose and 20 have red nose
- Protected attribute is nose color.
- Fairness: Statistical parity or Demographic parity - Equal outcome
  - Percentage of blue noses and red noses who get approved should be the same – evenly distribute resources
  - If 40 blue noses (50%) get their loan approved, there should be 10 for red noses (50%).
- Fairness: Equality of positive predictive value - Equal opportunity or accuracy
  - System prediction accuracy should be the same for each group - evenly distribute results
  - If 75% of bluenoses with approved loan are expected to pay back the loan, then it should also be 75% for red noses.
- Fairness: Equality of false negatives
  - Enforces constant false-negative rates across groups - evenly distribute “mistakes” made
  - If 25% of bluenoses who will pay back loan are wrongly rejected, then there should also be 25% of red noses
- Fairness: Through unawareness
  - Intentionally exclude protected attribute (nose color) and any related variables in decisions

Example: Fairness in Context

80 loan applicants, 20% will pay back

20 loan applicants, 40% will pay back



Protected Attribute:
Nose Color

20% (16)
Blue Nose

40% (8)
Red Nose

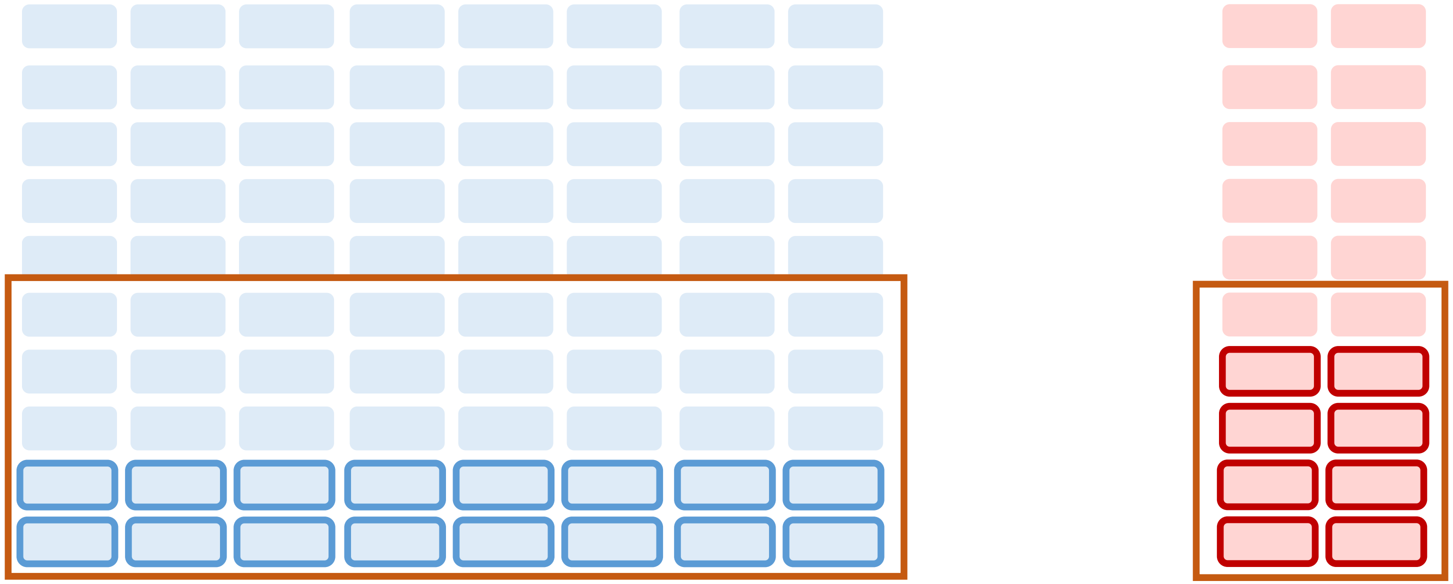
Example: Fairness in Context

- Consider 100 loan applications. 80 applicants have blue nose and 20 have red nose
- Protected attribute is nose color.
- Fairness: Statistical parity or Demographic parity - Equal outcome
  - Percentage of blue noses and red noses who get approved should be the same – evenly distribute resources
  - If 40 blue noses (50%) get their loan approved, there should be 10 for red noses (50%).

Example: Statistical or Demographic Parity

80 applicants, 20% will pay back

20 applicants, 40% will pay back



50% overall

Blue Nose

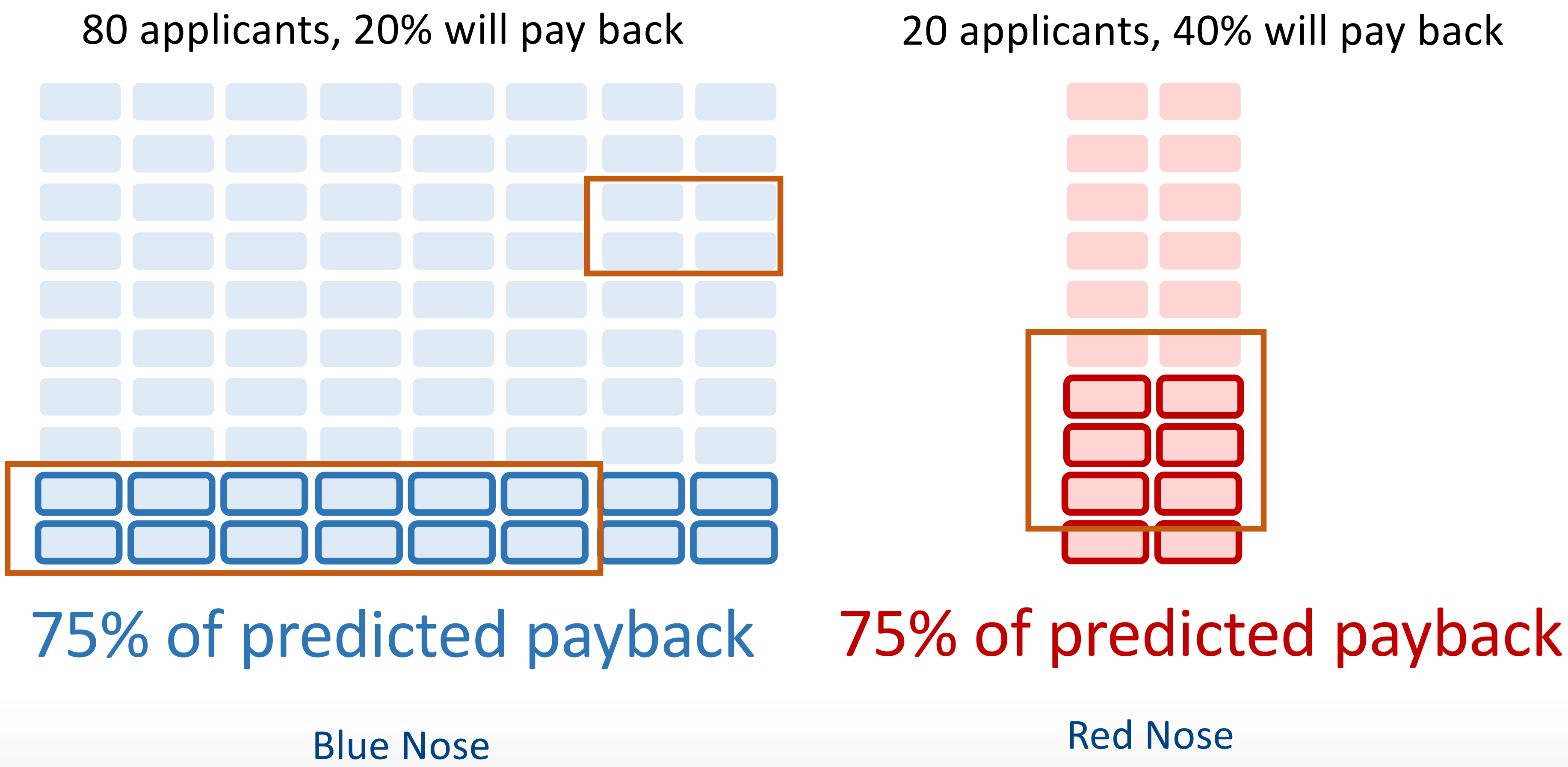
50% overall

Red Nose

Example: Fairness in Context

- Consider 100 loan applications. 80 applicants have blue nose and 20 have red nose
- Protected attribute is nose color.
- Fairness: Equality of positive predictive value - Equal opportunity or accuracy
  - System prediction accuracy should be the same for each group - evenly distribute results
  - If 75% of bluenoses with recommended loans are expected to pay back the loan, then it should also be 75% for red noses.

Example: Equality of Positive Predictive Value



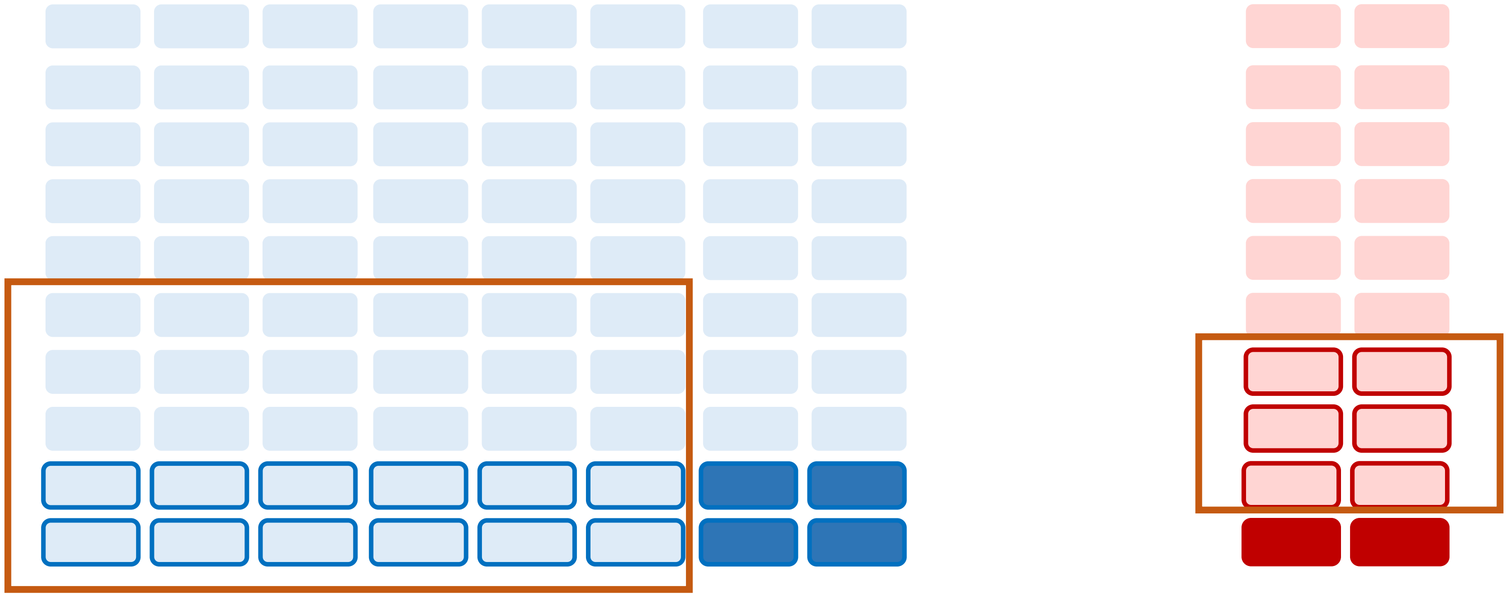
Example: Fairness in Context

- Consider 100 loan applications. 80 applicants have blue nose and 20 have red nose
- Protected attribute is nose color.
- Fairness: Equality of false negatives
  - Enforces constant false-negative rates across groups - evenly distribute “mistakes” made
  - If 25% of bluenoses who will pay back loan are wrongly rejected, then there should also be 25% of red noses

Example: Equality of False Negatives

80 applicants, 20% will pay back

20 applicants, 40% will pay back



25% of Wrong Rejections

25% of Wrong Rejections

Blue Nose

Red Nose

Example: Fairness in Context

- Consider 100 loan applications. 80 applicants have blue nose and 20 have red nose
- Protected attribute is nose color.
- Fairness: Through unawareness
  - Intentionally exclude protected attribute (nose color) and any related variables in decisions

- Impossible to achieve in practice!

Fairness

- Fairness criteria (examples):
  - Individual fairness
  - Group fairness
  - Fairness through unawareness
  - Statistical or demographical parity - Equal outcome -
  - Equal opportunity or accuracy - balance
  - Equal Impact - same expected utility
- Complications and issues:
  - Many fairness criteria
  - Choice of protected groups
  - Different bias classes and base rates
  - Sample size disparity
  - Inherent biases in data
  - Lack of unbiased ground truth data
  - Hard to implement in context

Ref: AIMA4e, Chapter 27

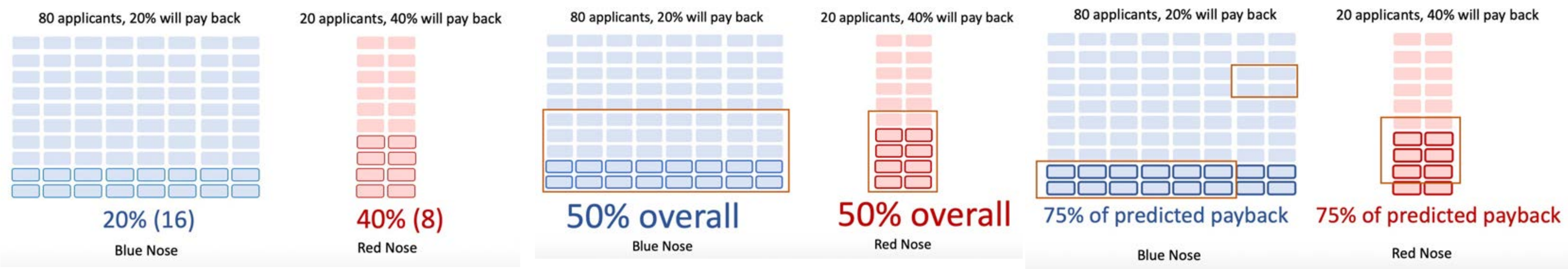
Ensuring Fairness and Mitigating Biases

- Methods to ensure fairness:
  - “Datasheets” for data
  - De-bias the data
  - Invent new bias-resistant algorithms
  - Two system approach - train second system to de-bias the recommendations of the first one
  - Informed human judgment and decision making
- Measurement and management of trade-offs:
  - Fairness vs Fairness
  - Fairness vs Accuracy

Ref: AIMA4e, Chapter 27

Subjective Judgment: Trade-off between Fairness Definitions

- Incompatible fairness definitions:
  - Combination of equality of both positive and negative rates across groups and
  - Equality of positive predictive value



Ref: The Ethical Algorithm, Chapter 2, 2019

No Fair Lunch: Pareto Frontier of Accuracy and Fairness

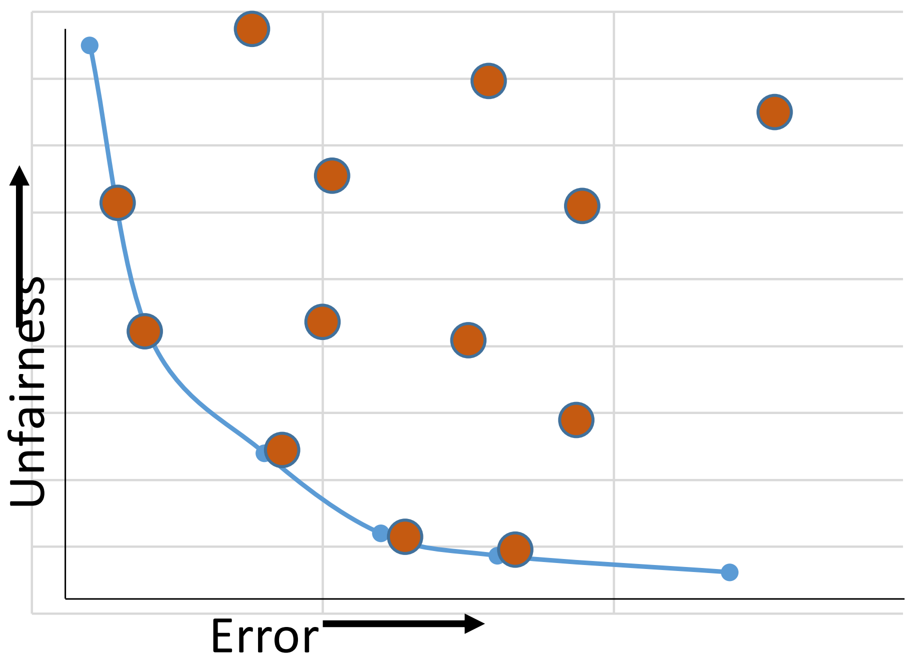
- ×��W !"#$%&'()*+,-./0123456789:;�ÿ�
    - #mistakes it makes on the data
    - “unfairness score” on the data

Pareto curve boundary:

- Captures “reasonable” choices for trade-off under fairness definition
- Quantifies “good” solutions to optimization problem with multiple competing criteria.
- Does not recommend “best” solution!

Human - chooses the one yielding the “best” trade-off or lowest overall error

Pareto Frontier: Accuracy-Fairness Trade-off



Ref: The Ethical Algorithm, Chapter 2, 2019

Making Fair Decisions

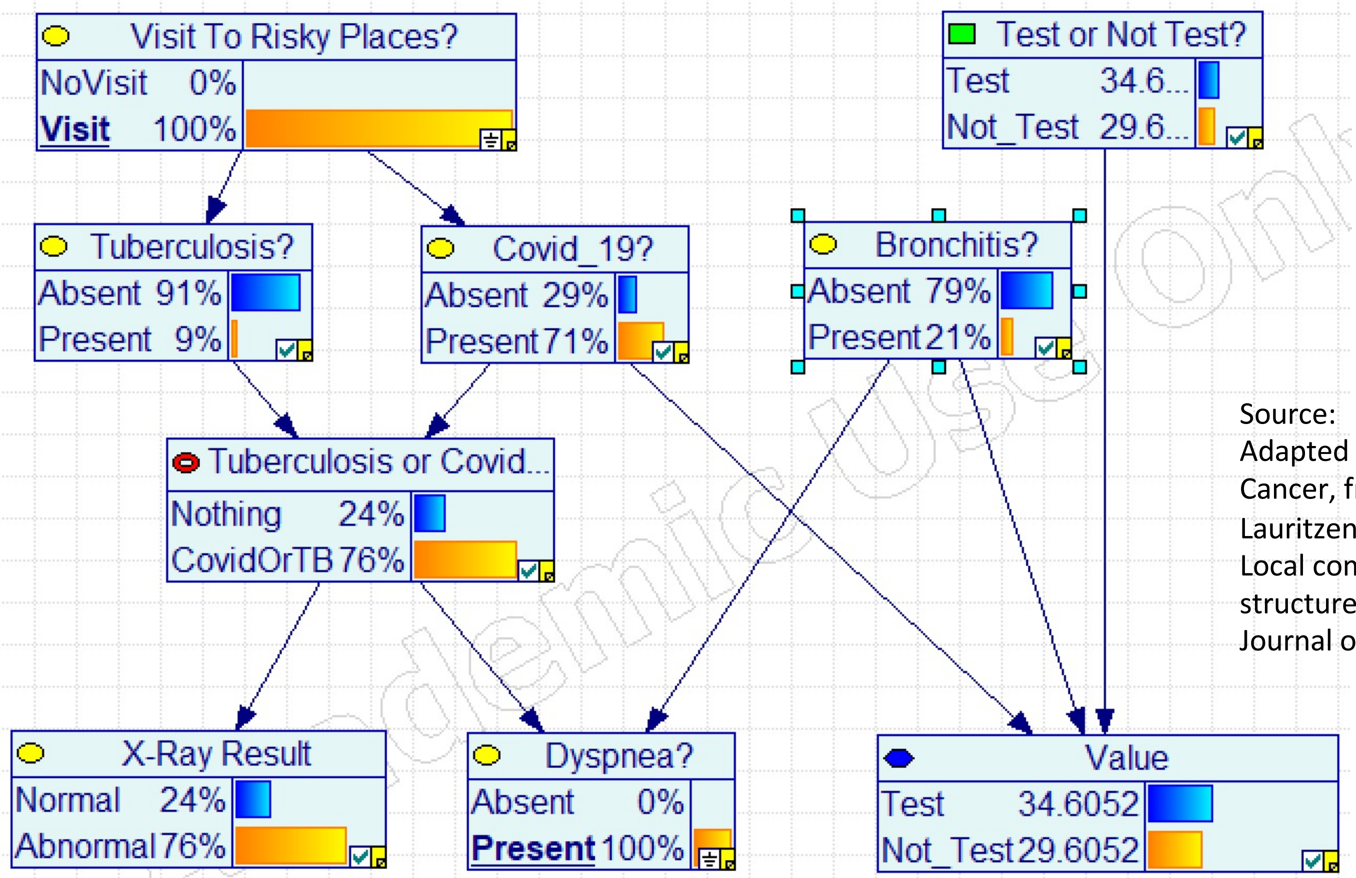
- Main issue:
  - Value gained from accurate insights balanced against individual or group fairness
- Examples
  - recruitment interviews, financial assistance schemes, university admissions
- If you are a developer, user, or owner/manager/regulator.
  1. What is fairness?
  2. What can be done to ensure fairness?
  3. What is the tradeoff between accuracy and fairness
  4. What are the implications?
  5. Who should make the decisions? When?

Myth and Mystery of XAI – Explainable AI

- A good explanation:
    - Understandable and convincing to the users
    - Accurately reflect the reasoning of the system
    - Complete and unambiguous
    - Specific for different users with different conditions or different outcomes
- Methods for XAI
    - Design AI algorithms with access to own deliberation processes – records as data structures
    - Certify machine explanations are not deceptions
    - Develop separate, aligned explanation system
    - Explanation + audit of past decisions, with aggregated statistics
    - Explanation may lead to revisions and remodeling – sf. Sensitivity analysis

Ref: AIMA4e, Chapter 27

Example: Explain yourself, AI …



Source:
Adapted from GENIE academic version Example – Lung Cancer, from the original network first appeared in: Lauritzen, Steffen L. & Spiegelhalter, David J. (1988). Local computations with probabilities on graphical structures and their application to expert systems, Journal of the Royal Statistical Society B, 50(2):157-224

Transparency

- Main issues:
  - What is going on inside a system? How to ensure correctness?
  - How to detect intentional malice, unintentional bug, or pervasive bias?
  - What about implicit intentions?
  - How to protect IP vs need for transparency for proper V&V or certification?
- Interpretable
  - Can we inspect content of the AI and see what it is doing?
  - Some methods for ML - SHapely Additive exPlanations (SHAP), LIME, DeepLIFT
- Explainable
  - Can we make up a "story" about what an AI is doing?
  - Explanation is helpful but not sufficient for trust
  - Explanations are convincing narratives about decisions

Ref: AIMA4e, Chapter 27; Lundberg and Lee, NIPS 2017

Making Transparent Decisions

- Main issue:
  - Value gained from results balanced against added justifications and clarifications
- Example – diagnosis recommendations, risk predictions
- If you are a developer, user, or owner/manager/regulator.
  1. What is transparency?
  2. What can be done to ensure transparency?
  3. What is the tradeoff between accuracy and accountability, responsibility, and transparency (ART) of a trusted agent?
  4. What are the implications?
  5. Who should make the decisions? When?

From Principles to Practice

- Hot topics of Ethical, Governance, Regulatory Considerations for AI
  - ~100+ public–private initiatives with high-level guidelines and principles for ethical development, deployment and governance of AI
  - Little practical deployment; processes and procedures unclear
- Challenge:
  - How to translate guidelines and principles into design requirements and technical components, governance frameworks and professional codes?
- Potential answer:
  - Pursue ethics (and governance and regulation) for AI as a process, not technological solutionism – in design, by design, for design(er)

Ref: Mittlestadt 2019, Vakkuri et. al. 2020, Dignum, V. 2019

Main Lessons

1. Building responsible AI requires close examination of ethical, governance, and regulatory considerations and decisions
2. Measurement and management of trade-offs between accuracy and relevant considerations to reach “good” decisions
3. Scientific methods and formal systems on the relevant topics are active research areas; some are ready for practical deployment
4. Human decision and judgment should work with these techniques at all stages of decision and policy making process
5. Take home ideas:
   - Human-aware approaches to developing responsible AI
   - Combination of human policies and scientific methods

Homework:

- Readings:
  - RN: 16.3.4 (Human judgment and irrationality)
  - RN: Chapter 27 (AI and ethics)
  - *(Optional)* Ong, D. An Ethical Framework for Guiding the Development of Affectively-Aware Artificial Intelligence. In: Proceedings of IEEE Affective Computing and Intelligent Interaction, 2021. https://arxiv.org/abs/2107.13734v1
- References on psychological traps and cognitive biases:

  (Journal articles publicly available online or through NUS Library e-Resources)
  - Tversky, A. and D. Kahneman, Judgment under Uncertainty: Heuristics and Biases. Science, 1974. 185(4157): p. 1124-1131
  - Kahneman, D. and A. Tversky, Prospect Theory: An Analysis of Decision under Risk. Econometrica, 1979. 47(2): p. 263-291.
  - Hammond, J.S., R.L. Keeney, and H. Raiffa, The hidden traps in decision making. Harvard Business Review, 1998. 76: p. 47+.
- General readings on cognitive biases:
  - Hammond, J.S., R.L. Keeney, and H. Raiffa, Smart Choices: A Practical Guide to Making Better Decisions. 2015, Harvard Business Review Press.
  - Kahneman, D., Thinking, fast and slow. Thinking, fast and slow. 2011, New York, NY, US: Farrar, Straus and Giroux.
  - Bazerman, M.H. and D.A. Moore, Judgment in Managerial Decision Making, 8th Edition. 2013: John Wiley & Sons.

Homework

- Background research on AI and Ethics:
  - **Covid-19 Contact Tracing Software**
  - What AI capabilities can or should be incorporated into the design?
  - What are the ethical, governance, and regulatory considerations?
  - What are your responsibilities if you are a developer, user, policy maker?
  - WHO policy brief on proximity tracking technologies:
    - https://apps.who.int/iris/bitstream/handle/10665/332200/WHO-2019-nCoV-Ethics_Contact_tracing_apps-2020.1-eng.pdf
  - WHO guidance on ethics and governance of AI for health:
    - https://www.who.int/publications/i/item/9789240029200
- References on AI and ethics:

  - Dietterich, T.G., Steps Toward Robust Artificial Intelligence. AI Magazine, 2017. 38(3).
  - Dignum, V., Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way. Artificial Intelligence: Foundations, Theory, and Algorithms 2019, Cham: Springer International Publishing.
  - Kearns, M. and A. Roth, The Ethical Algorithm: The Science of Socially Aware Algorithm Design. 2020: Oxford University Press. (Chapter 2)
  - Lundberg, S.M. and S.-I. Lee, A unified approach to interpreting model predictions, in Proceedings of the 31st International Conference on Neural Information Processing Systems. 2017, Curran Associates Inc.: Long Beach, California, USA. p. 4768–4777.
  - Mittelstadt, B., Principles alone cannot guarantee ethical AI. Nature Machine Intelligence, 2019. 1(11): p. 501-507.
  - Sculley, D., et al., Hidden technical debt in Machine learning systems, in Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2. 2015, MIT Press: Montreal, Canada. p. 2503–2511.
  - Vakkuri, V., et al., The Current State of Industrial Practice in Artificial Intelligence Ethics. IEEE Software, 2020. 37(4): p. 50-57.