14

Value Alignment

The first two chapters in this part of the book on interaction were focused on the communication from the machine system to the human consumer. This chapter is focused on the other direction of interaction: from humans to the machine system. Imagine that you’re the director of the selection committee of Alma Meadow, a (fictional) philanthropic organization that invests in early-stage social enterprises and invites the founders of those mission-driven organizations to participate in a two-year fellowship program. Alma Meadow receives about three thousand applications per year and selects about thirty of them to be fellowship recipients. As the director of this process, you are considering using machine learning in some capacity to improve the way it works. As such, you are a problem owner in the problem specification phase of an incipient machine learning lifecycle. Your main concern is that you do not sacrifice Alma Meadow’s mission or values in selecting social impact startups.

> “We need to have more conversations where we're doing this translation between policy, world outcome impact, what we care about and then all the math and data and tech stuff is in the back end trying to achieve these things.”
>
> —Rayid Ghani, machine learning and public policy researcher at Carnegie Mellon University

Values are fundamental beliefs that guide actions. They indicate the importance of various things and actions to a person or group of people, and determine the best ways to live and behave. Embedding Alma Meadow’s values in the machine learning system that you are contemplating is known as *value alignment* and has two parts.[1] The first part is *technical*: how to encode and elicit values in such a way that machine learning systems can access them and behave accordingly. The second part is *normative*: what the actual values are. (The word normative refers to norms in the social rather than mathematical sense: standards

[1] Iason Gabriel. “Artificial Intelligence, Values, and Alignment.” In: *Minds and Machines* 30 (Oct. 2020), pp. 411–437.

or principles of right action.) The focus of this chapter is on the first part of value alignment: the technical aspects for you, your colleagues, and other stakeholders to communicate your values (likely influenced by laws and regulations). The chapters in the sixth and final part of the book on purpose delve into the values themselves.

> “There is scientific research that can be undertaken to actually understand how to go from these values that we all agree on to embedding them into the AI system that’s working with humans.”
>
> —Francesca Rossi, AI ethics global leader at IBM

Before diving into the technical details of value alignment, let’s first take a step back and talk about two ways of expressing values: (1) deontological and (2) consequentionalist.[2] At a simplified level, *deontological* values are about defining good *actions* without concern for their outcomes, and *consequentialist* values are focused on defining *outcomes* that are good for all people. As an example, Alma Meadow has two deontological values: at least one of the recipients of the fellowship per year will be a formerly incarcerated individual and fellowship recipients’ social change organizations cannot promote a specific religious faith. These explicit rules or constraints on the action of awarding fellowships do not look into the effect on any outcome. In contrast, one of Alma Meadow’s consequentionalist values is that a fellowship recipient chosen from the applicant pool leads a social impact startup that will most improve the worldwide disability-adjusted life-years (DALY) in the next ten years. DALY is a metric that indicates the combined morbidity and mortality of the global disease burden. (It cannot be perfectly known which applicant satisfies this at the time the decision is made due to uncertainty, but it can still be a value.) It is a consequentionalist value because it is in terms of an outcome (DALY).

There is some overlap between deontology and procedural justice (described in Chapter 10), and between consequentionalism and distributive justice. One important difference between consequentialism and distributive justice is that in operationalizing distributive justice through group fairness as done in Chapter 10, the population over whom good outcomes are sought are the affected users, and that the justice/fairness is limited in time and scope to just the decision itself.[3] In contrast, in consequentionalism, the good is for all people throughout the broader society and the outcomes of interest are not only the immediate ones, but the longer term ones as well. Just like distributive justice was the focus in Chapter 10 rather than procedural justice because of its more natural operationalization in supervised classification, consequentialism is the focus here rather than deontology. However, it should be noted that deontological values may be elicited from people as rules and used as additional constraints to the Alma Meadow applicant screening model. In certain situations, such constraints can be easily added to the model without retraining.[4]

[2] Joshua Greene, Francesca Rossi, John Tasioulas, Kristen Brent Venable, and Brian Williams. “Embedding Ethical Principles in Collective Decision Support Systems.” In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Phoenix, Arizona, USA, Feb. 2016, pp. 4147–4151.

[3] Dallas Card and Noah A. Smith. “On Consequentialism and Fairness.” In: *Frontiers in Artificial Intelligence* 3.34 (May 2020).

[4] Elizabeth M. Daly, Massimiliano Mattetti, Öznur Alkan, and Rahul Nair. “User Driven Model Adjustment via Boolean Rule Explanations.” In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Feb. 2021, pp. 5896–5904.

It is critical not to take any shortcuts in value alignment because it forms the foundation for the other parts of the lifecycle. By going through the value alignment process, you arrive at problem specifications that data scientists try to satisfy using machine learning models, bias mitigation algorithms, explainability algorithms, adversarial defenses, etc. during the modeling phase of the lifecycle.

One thing to be wary of is underspecification that allows machine learning models to take shortcuts (also known as *specification gaming* and *reward hacking* in the value alignment literature).[5] This concept was covered in detail in Chapter 9, but is worth repeating. Any values that are left unsaid are free dimensions for machine learning algorithms to use as they please. So for example, even if the values you provide to the machine don't prioritize fairness, you might still be opposed to an extremely extremely unfair model in spirit. If you don't include at least some specification for a minimal level of fairness, the model may very well learn to be extremely unfair if it helps achieve specified values in accuracy, uncertainty quantification, and privacy.

In the remainder of the chapter, you will go through the problem specification phase for selecting Alma Meadow's fellows using supervised machine learning, insisting on value alignment. By the end, you'll have a better handle on the following questions.

- What are the different levels of consequentionalist values that you should consider?
- How should these values be elicited from individual people and fused together when elicited from a group of people?
- How do you put together elicited values with transparent documentation covered in Chapter 13 to *govern* machine learning systems?

14.1 Four Levels of Values in Trustworthy Machine Learning

When you were first starting to think about improving Alma Meadow's process for winnowing and selecting applications using machine learning, you had some rough idea why you wanted to do it (improving efficiency and transparency). However, you didn't have a progression of questions to work through as you figured out whether and in which parts of the selection process you should use machine learning, which pillars of trustworthy machine learning you should worry about, and how to make your worries quantitative. Let's list a series of four questions to help you gain clarity. (You'll be aided in answering them in the next section.)

1. Should you work on this problem?
2. Which pillars of trustworthiness are of concern?
3. What are the appropriate metrics for those pillars of trustworthiness?
4. What are acceptable ranges of the metric values?

The first question you should ask is whether you should even work on a problem. The answer may be no. If you stop and think for a minute, many problems are not problems to be solved. At face value,

[5] Victoria Krakovna, Jonathan Uesato, Vladimir Mikulik, Matthew Rahtz, Tom Everitt, Ramana Kumar, Zac Kenton, Jan Leike, and Shane Legg. "Specification Gaming: The Flip Side of AI Ingenuity." In: *DeepMind Blog* (Apr. 2020). URL: https://deepmind.com/blog/article/Specification-gaming-the-flip-side-of-AI-ingenuity.

evaluating three thousand applications and awarding fellowships seems not to be oppressive, harmful, misguided, or useless, but nevertheless, you should think deeply before answering.

> "Technical audiences are never satisfied with the fix being 'just don't do it.'"
>
> —Kristian Lum, statistician at University of Pennsylvania

Even if a problem is one that should be solved, machine learning is not always the answer. Alma Meadow has used a manual process to sort through applications for over thirty years, and has not been worse for wear. So why make the change now? Are there only some parts of the overall evaluation process for which machine learning makes sense?

The second question is more detailed. Among the different aspects of trustworthiness covered in the book so far, such as privacy, consent, accuracy, distributional robustness, fairness, adversarial robustness, interpretability, and uncertainty quantification, which ones are of the greatest concern? Are some essential and others only nice-to-haves? The third question takes the high-level elements of trustworthiness and brings them down to the level of specific metrics. Is accuracy, balanced accuracy, or AUC a more appropriate metric? How about the choice between statistical parity difference and average absolute odds difference? Lastly, the fourth question focuses on the preferred ranges of values of the metrics selected in the third question. Is a Brier score less than or equal to 0.25 acceptable? Importantly, there are relationships among the different pillars; you cannot create a system that is perfect in all respects. For example, typical differential privacy methods worsen fairness and uncertainty quantification.[6] Explainability may be at odds with other dimensions of trustworthiness.[7] Thus in the fourth question, it is critical to understand the relationships among metrics of different pillars and only specify ranges that are feasible.

14.2 *Representing and Eliciting Values*

Now that you have an overview of the four different levels of values for the supervised machine learning system you're contemplating for Alma Meadow's evaluation process, let's dig a little bit deeper to understand how to represent those values and how to make it easier for you to figure out what your values are.

14.2.1 *Should You Work on This Problem?*

A helpful tool in determining your values is a checklist of possible concerns along with case studies illustrating each of these concerns in real-world applications of machine learning related to your task of evaluating applications. An example of such a checklist and related case studies is the Ethical OS

[6] Marlotte Pannekoek and Giacomo Spigler. "Investigating Trade-Offs in Utility, Fairness and Differential Privacy in Neural Networks." arXiv:2102.05975, 2021. Zhiqi Bu, Hua Wang, Qi Long, and Weijie J. Su. "On the Convergence of Deep Learning with Differential Privacy." arXiv:2106.07830, 2021.

[7] Adrian Weller. "Transparency: Motivations and Challenges." In: *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. Ed. by Wojciech Samek, Grégoire Montavon, Andrea Vedaldi, Lars Kai Hansen, and Klaus-Robert Müller. Cham, Switzerland: Springer, 2019, pp. 23–40.

Toolkit,[8] which lists eight different broad consequences of the machine learning system that you should ponder:

1. Disinformation: the system helps subvert the truth at a large scale.
2. Addiction: the system keeps users engaged with it beyond what is good for them.
3. Economic inequality: the system contributes to income and wealth inequity by serving only well-heeled users or by eliminating low-income jobs.
4. Algorithmic bias: the system amplifies social biases.
5. Surveillance state: the system enables repression of dissent.
6. Loss of data control: the system causes people to lose control of their own personal data and any monetization it might lead to.
7. Surreptitious: the system does things that users don't know about.
8. Hate and crime: the system makes bullying, stalking, fraud, or theft easier.

Links to case studies accompany each of these checklist items in the Ethical OS Toolkit. Some of the case studies show when the item has happened in the real-world, and some show actions taken to prevent such items from happening. Another source of case studies is the continually-updated AI Incident Database.[9] Part 6 of the book, which is focused on purpose, touches on some of the items and case studies as well.

Starting with the checklist, your first step is to decide which items are good and which items are bad. In practice, you will read through the case studies, compare them to the Alma Meadow use case, spend some time thinking, and come up with your judgement. Many people, including you, will mark each of the eight items as bad, and judge the overall system to be too bad to proceed if any of them is true. But values are not universal. Some people may mark some of the checklist items as good. Some judgements may even be conditional. For example, with all else being equal, you might believe that algorithmic bias (item 4) is good if economic inequality (item 3) is false. In this second case and in even more complicated cases, reasoning about your preferences is not so easy.

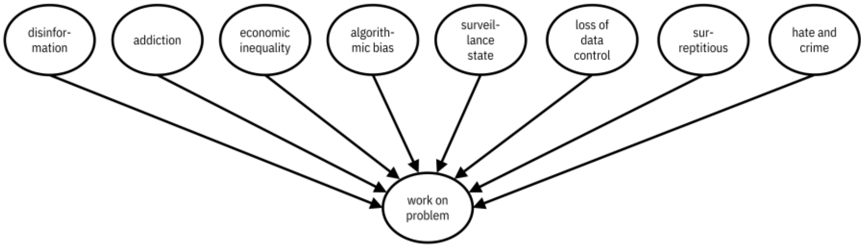
CP-nets are a representation of values, including conditional ones, that help you figure out your overall preference for the system and communicate it to the machine.[10] (The 'CP' stands for 'conditional preference.') CP-nets are directed graphical models with each node representing one attribute (checklist item) and arrows indicating conditional relationships. Each node also has a *conditional preference table* that gives the preferred values. (In this way, they are similar to causal graphs and structural equations you learned about in Chapter 8.) The symbol $\succ$ represents a preference relation; the argument on the left is preferred to the one on the right. The CP-net of the first case above (each of the eight items is bad) is given in Figure 14.1. It has an additional node at the bottom capturing the overall preference for working on the problem, which is conditioned on the eight items. There is a simple, greedy algorithm

[8] URL: https://ethicalos.org/wp-content/uploads/2018/08/Ethical-OS-Toolkit-2.pdf

[9] Sean McGregor. "Preventing Repeated Real World AI Failures by Cataloging Incidents: The AI Incident Database." In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Feb. 2021, pp. 15458–15463.

[10] Craig Boutilier, Ronen I. Brafman, Carmel Domshlak, Holger H. Hoos, and David Poole. "CP-Nets: A Tool for Representing and Reasoning with Conditional Ceteris Paribus Preference Statements." In: *Journal of Artificial Intelligence Research* 21.1 (Jan. 2004), pp. 135–191.

for figuring out the most preferred instantiation of the values from CP-nets. However, in this case it is easy to figure out the answer without an algorithm: it is the system that does not satisfy any of the eight checklist items and says to go ahead and work on the problem. In more general cases with complicated CP-nets, the inference algorithm is helpful.



disinformation	addiction	economic inequity	algorithmic bias	surveillance state	loss of data control	surreptitious	hate and crime
no ≻ yes	no ≻ yes	no ≻ yes	no ≻ yes	no ≻ yes	no ≻ yes	no ≻ yes	no ≻ yes

work on problem
disinformation = no, addiction = no, economic inequity = no, algorithmic bias = no, surveillance state = no, loss of data control = no, surreptitious = no, hate and crime = no: yes ≻ no
disinformation = yes, addiction = no, economic inequity = no, algorithmic bias = no, surveillance state = no, loss of data control = no, surreptitious = no, hate and crime = no: no ≻ yes
disinformation = no, addiction = yes, economic inequity = no, algorithmic bias = no, surveillance state = no, loss of data control = no, surreptitious = no, hate and crime = no: no ≻ yes
⋮
disinformation = no, addiction = no, economic inequity = no, algorithmic bias = no, surveillance state = no, loss of data control = no, surreptitious = no, hate and crime = yes: no ≻ yes
disinformation = yes, addiction = yes, economic inequity = no, algorithmic bias = no, surveillance state = no, loss of data control = no, surreptitious = no, hate and crime = no: no ≻ yes
⋮
disinformation = yes, addiction = yes, economic inequity = yes, algorithmic bias = yes, surveillance state = yes, loss of data control = yes, surreptitious = yes, hate and crime = yes: no ≻ yes

Figure 14.1. *An example CP-net for whether Alma Meadow should work on the application evaluation problem. At the top is the graphical model. At the bottom are the conditional preference tables.* Accessible caption. Eight nodes disinformation, addiction, economic inequity, algorithmic bias, surveillance state, loss of data control, surreptitious, and hate and crime all have the node work on problem as their child. All preferences for the top eight nodes are no ≻ yes. In all configurations of yeses and noes, the work on problem preference is no ≻ yes, except when all top eight nodes have configuration no, when it is no ≻ yes.

With the values decided, it is time to go through the checklist items and determine whether they are consistent with your most preferred values:

1. Disinformation = no: evaluating applications from social entrepreneurs is unlikely to subvert the truth.
2. Addiction = no: this use of machine learning is not likely to lead to addiction.
3. Economic inequality = partly yes, partly no: it is possible the system could only select applications that have very technical descriptions of the social impact startup's value proposition and have been professionally polished. However, this possibility is not enough of a concern to completely stop the use of machine learning. What this concern does suggest, though, is that machine learning only be used to prioritize semi-finalists rather than later in the evaluation process because human evaluators may find gems that seem unusual to the machine.

4. Algorithmic bias = no: Alma Meadow has been extremely proactive in preventing social bias with respect to common protected attributes in its human evaluations in past years, so the training data will not yield much social bias in models.
5. Surveillance state = no: the machine learning system is unlikely to be an instrument of oppression.
6. Loss of data control = no: by sharing their ideas in the application, budding social entrepreneurs could feel that they are giving up their intellectual property, but Alma Meadow has gone to great lengths to ensure that is not the case. In fact, toward one of its values, Alma Meadow provides information to applicants on how to construct confidential information assignment agreements.
7. Surreptitious = no: the system is unlikely to do anything users don't know about.
8. Hate and crime = no: the system is unlikely to enable criminal activities.

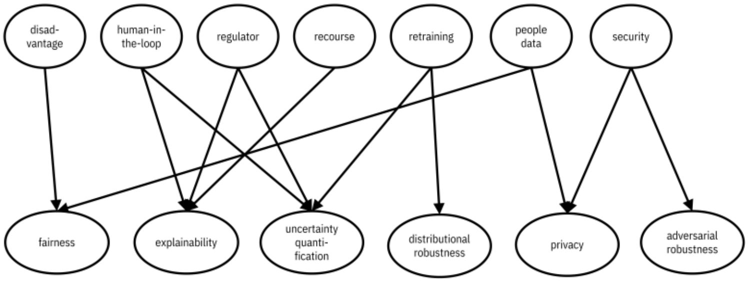
None of the items are properties of the system, including economic inequality when restricting the use of machine learning only to a first-round prioritization. This is consistent with your most-preferred values, so you should work on this problem.

14.2.2 Which Pillars of Trustworthiness Are of Concern?

Now that you have passed the first level of value judgement, you have to determine which elements of trust are your top priority in the feature engineering and modeling phases. Rather than having you take on the very difficult task of trying to directly state a preference ordering, e.g. fairness $\succ$ explainability $\succ$ distributional robustness $\succ$ uncertainty quantification $\succ$ privacy $\succ$ adversarial robustness, let's create a CP-net with some considerations that are easier to answer. To make things even easier, let's assume that you are in a predictive modeling situation, not causal modeling of interventions. Let's take accuracy and similar performance metrics from Chapter 6 out of the equation, since basic competence is always valued. Furthermore, assume the application is high-risk (true for Alma Meadow's applicant selection), so the different elements of trustworthiness are part of your value consideration, and assume that consent and transparency are required. Then a construction of the CP-net for pillars of trustworthiness begins with the following seven properties:

1. Disadvantage (no, yes): the decisions have the possibility of giving systematic disadvantage to certain groups or individuals.
2. Human-in-the-loop (no, yes): the system predictions support a human decision-maker.
3. Regulator (no, yes): regulators (broadly-construed) audit the model.
4. Recourse (no, yes): affected users of the system have the ability to challenge the decision they receive.
5. Retraining (no, yes): the model is retrained frequently to match the time scale of distribution shift.
6. People data (not about people, about people but not SPI, SPI): the system may use data about people which may be sensitive personal information (SPI).
7. Security (external, internal and not secure, secure): the data, model interface, or software code are available either externally or only internally, and may be kept highly secured.

Once you have given these seven system preferences, giving conditional preferences for the different elements of trustworthiness is more compact. They can simply be given as high or low priority values based on just a few of the system preferences. For example, if there is a possibility of systematic disadvantage *and* the problem involves people data, then giving attention to fairness may be highly valued. Putting everything together yields a CP-net like the one in Figure 14.2.



disadvantage
yes ≻ no

human-in-the-loop
yes ≻ no

regulator
no ≻ yes

recourse
yes ≻ no

retraining
no ≻ yes

people data
not SPI ≻ not about people ≻ SPI

security
secure ≻ internal and not secure ≻ external

fairness
disadvantage = no AND people data = not about people: low priority ≻ high priority
disadvantage = no AND people data = not SPI: low priority ≻ high priority
disadvantage = no AND people data = SPI: low priority ≻ high priority
disadvantage = yes AND people data = not about people: low priority ≻ high priority
disadvantage = yes AND people data = not SPI: high priority ≻ low priority
disadvantage = yes AND people data = SPI: high priority ≻ low priority

distributional robustness
retraining = no: high priority ≻ low priority
retraining = yes: low priority ≻ high priority

explainability
human-in-the-loop = no AND regulator = no AND recourse = no: low priority ≻ high priority
human-in-the-loop = no AND regulator = no AND recourse = yes: high priority ≻ low priority
human-in-the-loop = no AND regulator = yes AND recourse = no: high priority ≻ low priority
human-in-the-loop = no AND regulator = yes AND recourse = yes: high priority ≻ low priority
human-in-the-loop = yes AND regulator = no AND recourse = no: high priority ≻ low priority
human-in-the-loop = yes AND regulator = no AND recourse = yes: high priority ≻ low priority
human-in-the-loop = yes AND regulator = yes AND recourse = no: high priority ≻ low priority
human-in-the-loop = yes AND regulator = yes AND recourse = yes: high priority ≻ low priority

uncertainty quantification
human-in-the-loop = no AND regulator = no AND retraining = no: high priority ≻ low priority
human-in-the-loop = no AND regulator = no AND retraining = yes: low priority ≻ high priority
human-in-the-loop = no AND regulator = yes AND retraining = no: high priority ≻ low priority
human-in-the-loop = no AND regulator = yes AND retraining = yes: high priority ≻ low priority
human-in-the-loop = yes AND regulator = no AND retraining = no: high priority ≻ low priority
human-in-the-loop = yes AND regulator = no AND retraining = yes: high priority ≻ low priority
human-in-the-loop = yes AND regulator = yes AND retraining = no: high priority ≻ low priority
human-in-the-loop = yes AND regulator = yes AND retraining = yes: high priority ≻ low priority

privacy
people data = not about people AND security = external: low priority ≻ high priority
people data = not about people AND security = internal and not secure: low priority ≻ high priority
people data = not about people AND security = secure: low priority ≻ high priority
people data = not SPI AND security = external: low priority ≻ high priority
people data = not SPI AND security = internal and not secure: low priority ≻ high priority
people data = not SPI AND security = secure: low priority ≻ high priority
people data = SPI AND security = external: high priority ≻ low priority
people data = SPI AND security = internal and not secure: low priority ≻ high priority
people data = SPI AND security = secure: low priority ≻ high priority

adversarial robustness
security = external: high priority ≻ low priority
security = internal and not secure: high priority ≻ low priority
security = secure: low priority ≻ high priority

Figure 14.2. *An example CP-net for which pillars of trustworthiness Alma Meadow should prioritize when developing a model for the application evaluation problem. At the top is the graphical model. At the bottom are the conditional preference tables.* Accessible caption. In the graphical model, there are edges from disadvantage to fairness, people data to fairness, human-in-the-loop to explainability, regulator to explainability, recourse to explainability, human-in-the-loop to uncertainty quantification, regulator to uncertainty quantification, retraining to uncertainty quantification, retraining to distributional robustness, people data to privacy, security to privacy, and security to adversarial robustness. The conditional preference tables list many different complicated preferences.

The top-level system property preferences will be highly specific to your Alma Meadow application evaluation use case. You and other problem owners have the requisite knowledge at your fingertips to provide your judgements. The conditional preferences connecting the top-level properties with the specific elements of trustworthiness (fairness, explainability, etc.) are more generic and generalizable. Even if the edges and conditional preference tables given in the figure are not 100% universal, they are close to universal and can be used as-is in many different application domains.

In the Alma Meadow example in Figure 14.2, your specific judgements are: systematic disadvantage is possible, you prefer a human decision-maker in the loop, there will not be a regulator audit, you prefer that social entrepreneur applicants have an opportunity for recourse, you prefer the system not be retrained frequently, you prefer that the applications contain data about people (both about the applicant and the population their organization serves) but not anything personally-sensitive, and you prefer that the data and models be secured. Based on these values and the conditional preferences lower in the CP-net, the following pillars are inferred to be higher priority: fairness, explainability, uncertainty quantification, and distributional robustness. Privacy and adversarial robustness are inferred to be lower priority.

14.2.3 What Are the Appropriate Metrics?

After the second stage of value alignment, you know which pillars of trustworthiness are higher priority and you can move on to figuring out specific metrics within the pillars. This problem is known as *performance metric elicitation*. In previous chapters, you've already learned about different considerations when making these determinations. For example, in Chapter 6, it was discussed that AUC is an appropriate basic performance metric when you desire good performance across all operating points. As another example, Table 10.1 summarized the considerations in determining group fairness metrics: whether you are testing data or models, whether there is social bias in the measurement process, and whether the favorable label is assistive or non-punitive. We will not repeat those arguments here, which you should definitely go through, but will mention another tool to help you in metric elicitation.

In the previous elicitation task, it was difficult to go straight to a total preference ordering for the different pillars of trustworthiness; the task was made easier by asking simpler and more structured judgements using CP-nets. There's a similar story here, but using *pairwise comparisons* instead of CP-nets. The elicitation process is like an optometrist helping you home in on your preferred eye prescription by having you compare a sequence of pairs of lenses. Here, the pairwise comparisons are between different possible metrics within a given pillar. By comparing the values of two metrics for many models, you get a sense of what they're indicating and can choose one over the other. If the pairs are chosen in an intelligent way and you do enough comparisons, you will converge onto your preferred metric. One such intelligent way efficiently elicits basic performance metrics and fairness metrics by taking advantage of their linearity or quadraticity properties and showing users a sequence of pairs of confusion matrices (recall confusion matrices from Chapter 6).[11] Confusion matrices may be too difficult for different stakeholders to reason about in their typical format as a 2×2 matrix of numbers; alternate visualizations of confusion matrices such as tree diagrams, flow charts, and matrices presented with

[11] Gaurush Hiranandani, Harikrishna Narasimhan, and Oluwasanmi Koyejo. "Fair Performance Metric Elicitation." In: *Advances in Neural Information Processing Systems* 33 (Dec. 2020), pp. 11083–11095.

contextual information may be used instead.[12] Another approach based on pairwise comparisons is known as the *analytical hierarchy process*; it asks for numerical ratings (one to nine) in the comparison so that you not only indicate which metric is better, but by roughly how much as well.[13]

14.2.4 What are Acceptable Ranges of the Metric Values?

Once specific metrics have been selected, the final level of value alignment is determining the quantitative ranges of preferred metric values for the Alma Meadow semi-finalist selection model. Since the different elements of trustworthiness and their relevant metrics are interrelated, including some that are tradeoffs, this level of elicitation should not be approached one metric at a time like the previous metric elicitation, but more holistically.

The starting point is a feasible set of metric values, shown schematically in Figure 14.3. In this schematic, the quantitative test results for a single model (shown as tables, bar graphs, parallel coordinate plots, and radar charts in Chapter 13) are mapped to a single point inside the feasible region. From Chapter 6, you know that the optimal Bayes risk is fundamentally the best you can ever do for cost-weighted accuracy. As also mentioned in that chapter, it turns out that you can empirically estimate the optimal Bayes risk from the historical Alma Meadow applications data you have.[14] Moreover, fundamental theoretical relationships between metrics from different elements of trustworthiness are starting to be researched using the concept of Chernoff information[15] from detection theory and information theory (they include both tradeoffs and non-tradeoffs): a so-called unified theory of trust.[16] Once that research is completed, the schematic diagram of Figure 14.3 can be actualized for a given machine learning task and the fourth value alignment question (ranges of values of different metrics) can be more easily stated. By explicitly knowing the feasible set of metric values, you can confidently make choices that are possible for the Alma Meadow semi-finalist prioritization model instead of wishful thinking.

[12] Hong Shen, Haojian Jin, Ángel Alexander Cabrera, Adam Perer, Haiyi Zhu, and Jason I. Hong. "Designing Alternative Representations of Confusion Matrices to Support Non-Expert Public Understanding of Algorithm Performance." In: *Proceedings of the ACM on Human-Computer Interaction* 4.CSCW2 (Oct. 2020), p. 153.

[13] Yunfeng Zhang, Rachel K. E. Bellamy, and Kush R. Varshney. "Joint Optimization of AI Fairness and Utility: A Human-Centered Approach." In: *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. New York, New York, USA, Feb. 2020, pp. 400–406.

[14] Visar Berisha, Alan Wisler, Alfred O. Hero, III, and Andreas Spanias. "Empirically Estimable Classification Bounds Based on a Nonparametric Divergence Measure." In: *IEEE Transactions on Signal Processing* 64.3 (Feb. 2016), pp. 580–591. Ryan Theisen, Huan Wang, Lav R. Varshney, Caiming Xiong, and Richard Socher. "Evaluating State-of-the-Art Classification Models Against Bayes Optimality." In: *Advances in Neural Processing Systems* 34 (Dec. 2021).

[15] Frank Nielsen. "An Information-Geometric Characterization of Chernoff Information." In: *IEEE Signal Processing Letters* 20.3 (Mar. 2013), pp. 269–272.

[16] Sanghamitra Dutta, Dennis Wei, Hazar Yueksel, Pin-Yu Chen, Sijia Liu, and Kush R. Varshney, "Is There a Trade-Off Between Fairness and Accuracy? A Perspective Using Mismatched Hypothesis Testing." In: *Proceedings of the International Conference on Machine Learning*. Jul. 2020, pp. 2803–2813. Kush R. Varshney, Prashant Khanduri, Pranay Sharma, Shan Zhang, and Pramod K. Varshney, "Why Interpretability in Machine Learning? An Answer Using Distributed Detection and Data Fusion Theory." In: *Proceedings of the ICML Workshop on Human Interpretability in Machine Learning*. Stockholm, Sweden, Jul. 2018, pp. 15–20. Zuxing Li, Tobias J. Oechtering, and Deniz Gündüz. "Privacy Against a Hypothesis Testing Adversary." In: *IEEE Transactions on Information Forensics and Security* 14.6 (Jun. 2019), pp. 1567–1581.

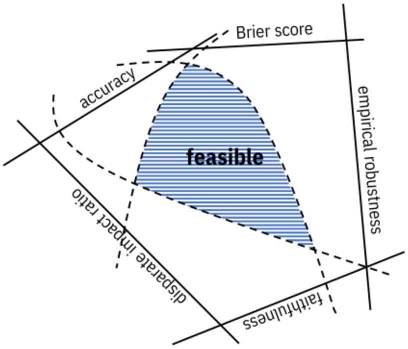


Figure 14.3. *Schematic diagram of feasible set of trust-related metrics.* Accessible caption. A shaded region enclosed by three curved segments is labeled feasible. It is surrounded by five axes: accuracy, Brier score, empirical robustness, faithfulness, and disparate impact ratio.

The feasible set is a good starting point, but there is still the question of deciding on the preferred ranges of the metrics. Two approaches may help. First, a value alignment system can automatically collect or create a corpus of many models for the same or similar prediction task and compute their metrics. This will yield an empirical characterization of the interrelationships among the metrics.[17] You can better understand your choice of metric values based on their joint distribution in the corpus. The joint distribution can be visualized using a parallel coordinate density plot mentioned in Chapter 13.

Second, the value alignment system can utilize a variation of so-called *trolley problems* for supervised machine learning. A trolley problem is a thought experiment about a fictional situation in which you can save the lives of five people who'll otherwise be hit by a trolley by swerving and killing one person. Whether you choose to divert the trolley reveals your values. Variations of trolley problems change the number of people who die under each option and associate attributes with the people.[18] They are also pairwise comparisons. Trolley problems are useful for value elicitation because humans are more easily able to reason about small numbers than the long decimals that usually appear in trust metrics. Moreover, couching judgements in terms of an actual scenario helps people internalize the consequences of the decision and relate them to their use case.

As an example, consider the two scenarios shown in Figure 14.4. Which one do you prefer? Would you rather have an adversarial example fool the system or have a large disparate impact ratio? The actual numbers also play a role because a disparate impact ratio of 2 in scenario 2 is quite high. There is no right or wrong answer, but whatever you select indicates your values.

[17] Moninder Singh, Gevorg Ghalachyan, Kush R. Varshney, and Reginald E. Bryant. "An Empirical Study of Accuracy, Fairness, Explainability, Distributional Robustness, and Adversarial Robustness." In: *KDD Workshop on Measures and Best Practices for Responsible AI*. Aug. 2021.

[18] Edmond Awad, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, Jean-François Bonnefon, and Iyad Rahwan. "The Moral Machine Experiment." In: *Nature* 563.7729 (Oct. 2018), pp. 59–64.

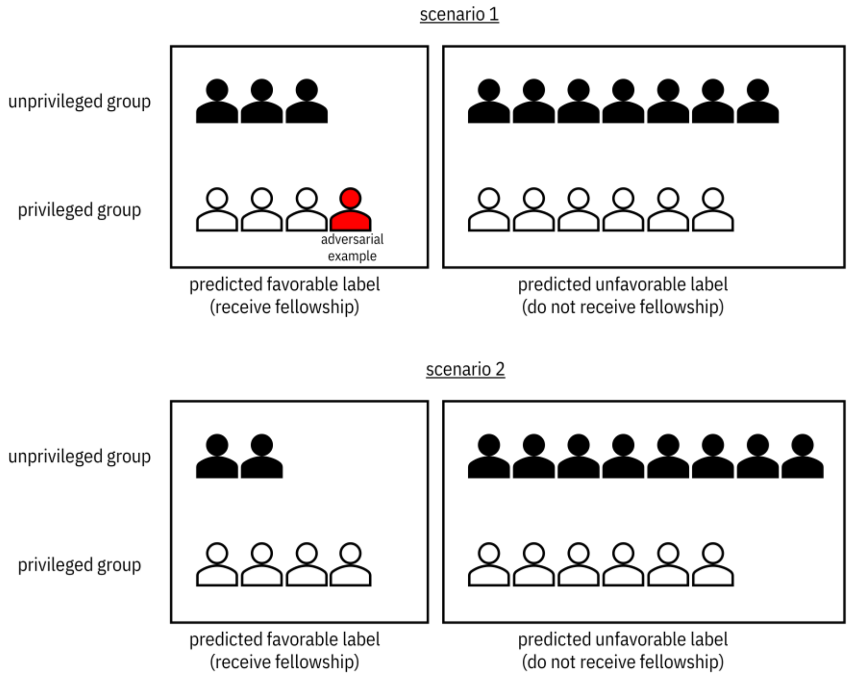

scenario 1
unprivileged group
privileged group
adversarial example
predicted favorable label (receive fellowship)
predicted unfavorable label (do not receive fellowship)
scenario 2
unprivileged group
privileged group
predicted favorable label (receive fellowship)
predicted unfavorable label (do not receive fellowship)

Figure 14.4. *A pairwise comparison of illustrated scenarios.* Accessible caption. Two scenarios each have different small numbers of members of unprivileged and privileged groups receiving and not receiving the fellowship. The first scenario also has an adversarial example.

14.3 Fusion of Preferences Over a Group

Based on the previous section, you have several ways to tell the machine learning system your preferred values at different levels of granularity. As the problem owner, you have a lot of power. But should you wield that power unilaterally? Wouldn't it better to include diverse voices and build consensus? Yes it would; it is important to take the preferences of other stakeholders such as the executive director, board members, and members of the Alma Meadow team into account. It is also critical that budding social entrepreneurs and the beneficiaries of their social impact startups participate in the value alignment process (they should be monetarily compensated for participating). The values communicated to the machine learning system should also take applicable laws and regulations into account; the law is another voice.

Each of the individuals in an assembled panel can go through the same four-level value elicitation that you did in the previous section, yielding several CP-nets and sets of pairwise comparisons. But then what? How do you technically combine the individual preferences expressed by the different folks? *Voting* of some kind, also known as *computational social choice*, is a natural answer. Extensions of both CP-nets and the analytic hierarchy process use voting-like mechanisms to fuse together several individual preferences.[19] Other methods for aggregating individual preferences into collective preferences are also based on voting.[20]

Voting methods typically aim to choose the value that is preferred by the majority in every pairwise comparison with other possible values (this majority-preferred set of values is known as the *Condorcet winner*). However, it is not clear if such majoritarianism is really what you want when combining the preferences of the various stakeholders. Minority voices may raise important points that shouldn't be drowned out by the majority, which is apt to happen in independent individual elicitation followed by a voting-based preference fusion. The degree of participation by members of minoritized groups should not be so weak as to be meaningless or even worse: extractive (the idea of extraction conceived in postcolonialialism is covered in Chapter 15).[21] This shortcoming of voting systems suggests that an alternative process be pursued that does not reproduce existing power dynamics. *Participatory design*—various stakeholders, data scientists and engineers working together in facilitated sessions to collectively come up with single CP-nets and pairwise comparisons—is a suggested remedy, but may in fact also reproduce existing power dynamics if not conducted well. So in your role at Alma Meadow, don't skimp on well-trained facilitators for participatory design sessions.

14.4 Governance

You've come to an agreement with the stakeholders on the values that should be expressed in Alma Meadow's application screening system. You've specified them as feasible ranges of quantitative metrics that the machine learning system can incorporate. Now how do you ensure that those desired values are realized by the deployed machine learning model? Through *control* or *governance*.[22] Viewing the lifecycle as a control system, illustrated in Figure 14.5, the values coming out of value alignment are the reference input, the data scientists are the controllers that try to do all they can so the machine learning system meets the desired values, and model facts (described in Chapter 13 as part of transparency) are the measured output of testing that indicate whether the values are met. Any difference between the facts and the values is a signal of misalignment to the data scientists; they must do a better job in modeling. In this way, the governance of machine learning systems requires both the elicitation of the system's desired behavior (value alignment) and the reporting of facts that measure those behaviors (transparency).

[19] Lirong Xia, Vincent Conitzer, and Jérôme Lang. "Voting on Multiattribute Domains with Cyclic Preferential Dependencies." In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Chicago, Illinois, USA, Jul. 2008, pp. 202–207. Indrani Basak and Thomas Saaty. "Group Decision Making Using the Analytic Hierarchy Process." In: *Mathematical and Computer Modelling* 17.4–5 (Feb.–Mar. 1993), pp. 101–109.

[20] Ritesh Noothigattu, Snehalkumar 'Neil' S. Gaikwad, Edmond Awad, Sohan Dsouza, Iyad Rahwan, Pradeep Ravikumar, and Ariel D. Procaccia. "A Voting-Based System for Ethical Decision Making." In: *Proceedings of the AAAI Conference on Artificial Intelligence*. New Orleans, Louisiana, USA, Feb. 2018, pp. 1587–1594. Min Kyung Lee, Daniel Kusbit, Anson Kahng, Ji Tae Kim, Xinran Yuan, Allissa Chan, Daniel See, Ritesh Noothigattu, Siheon Lee, Alexandros Psomas, and Ariel D. Procaccia. "WeBuildAI: Participatory Framework for Algorithmic Governance." In: *Proceedings of the ACM on Human-Computer Interaction* 3.181 (Nov. 2019).

[21] Sasha Costanza-Chock. *Design Justice: Community-Led Practices to Build the Worlds We Need*. Cambridge, Massachusetts, USA: MIT Press, 2020.

[22] Osonde A. Osoba, Benjamin Boudreaux, and Douglas Yeung. "Steps Towards Value-Aligned Systems." In: *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. New York, New York, USA, Feb. 2020, pp. 332–336.

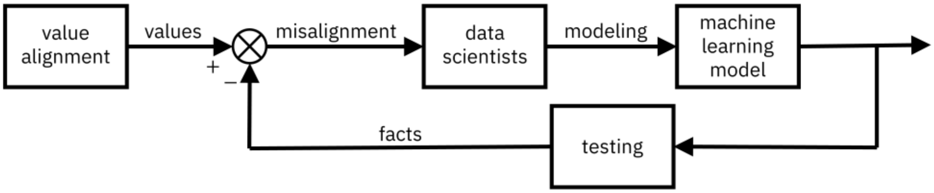


Figure 14.5. *Transparent documentation and value alignment come together to help in the governance of machine learning systems.* Accessible caption. A block diagram that starts with a value alignment block out of which come values. Facts are subtracted from values to yield misalignment. Misalignment is input to a data scientists block with modeling as output. Modeling is input to a machine learning model with output that is fed into a testing block. The output of testing is the same facts that were subtracted from values, creating a feedback loop.

In Chapter 13, factsheets contained not only quantitative test results, but also intended uses and other qualitative knowledge about the development process. However, in the view of governance presented here, only the quantitative test results seem to be used. So, is governance concerned only with test outcomes, which are of a consequentialist nature, or is it also concerned with the development process, which is of a deontological nature? Since the controllers—the data scientists—are people with inherent quirks and biases, both kinds of facts together help them see the big picture goals without losing track of their lower-level, day-to-day duties for resolving misalignment. Thus, a codification of processes to be followed during development is an integral part of governance. Toward this end, you have instituted a set of checklists for Alma Meadow's data scientists to follow, resulting in a well-governed system overall.

14.5 Summary

- Interaction between people and machine learning systems is not only *from* the machine learning system *to* a human via explainability and transparency. The other direction from humans to the machine, known as value alignment, is just as critical so that people can instruct the machine on acceptable behaviors.
- There are two kinds of values: consequentialist values that are concerned with outcomes and deontological values that are concerned with actions. Consequentialist values are more natural in value alignment for supervised machine learning systems.
- Value alignment for supervised classification consists of four levels. Should you work on a problem? Which pillars of trustworthiness are high priority? What are the appropriate metrics? What are acceptable metric value ranges?
- CP-nets and pairwise comparisons are tools for structuring the elicitation of preferences of values across the four levels.
- The preferences of a group of stakeholders, including those from traditionally marginalized backgrounds, may be combined using either voting or participatory design sessions.
- Governance of machine learning systems combines value alignment to elicit desired behaviors with factsheet-based transparency to measure whether those elicited behaviors are being met.