The *Situate AI* Guidebook

The *Situate AI* Guidebook is a process to scaffold **early-stage deliberations** around ***whether* to move forward with a new AI tool idea, design, or development**.

The guidebook provides resources–including **reflective prompts, response guidance, and a deliverable template**–to guide your organization through this deliberative decision-making process.

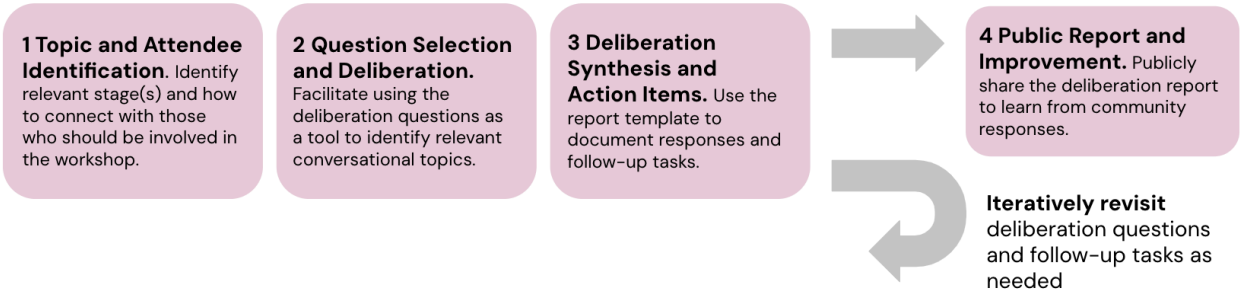
Overview

After completing the Situate AI Guidebook, your organization should have a recommendation for the prompt: **Should we move forward with implementing the AI tool? If yes, what are key considerations to plan for?**

This Situate AI Guidebook supports you in forming this recommendation through a deliberation-driven process involving:

1. **Question prompts to support conversations** around the social (organizational, societal, and legal) and technical (data and modeling) considerations that should provide supporting evidence for your recommendation.
2. **Pointers to external resources** to help guide your responses.
3. **Template for a recommended deliverable** to help formulate evidence for your recommendation based on the deliberations.

The Situate AI Guidebook process overview:



Link to recommendation deliverable report template.

Who is this guidebook for?

The current version of this guidebook is intended for use within **public sector agencies** at various stages of maturity in their use of AI tools–from those that are just beginning to consider the use of new AI tools to those that may already have years of experience implementing their own AI tools. The deliberation questions are designed to be discussed across different stakeholders employed in a public sector agency: Agency leadership, AI practitioners and analysts, program managers, and frontline workers. We recommend having one facilitator for any given deliberation workshop.

Which project stages does this guidebook target?

This guidebook includes branched questions and resources to support decision-making at the any of the following three project stages:

1. There is an **idea** for a potential new AI tool, but it has not yet been designed.
2. There is a **design** for a potential new AI tool, but it has not yet been developed.
3. A new AI tool has **already been developed**, but has not yet been implemented in practice.

How to use this guidebook

(section coming soon!)

Toolkit Questions

There are four facets of deliberative questions.

1. Goals and Intended Use
2. Societal and Legal Considerations
3. Data and Modeling Constraints
4. Organizational Governance Factors

Facet 1: Goals and Intended Use (55 questions)

The set of questions below are intended to support conversations around the following broader question: **Given our underlying goals and intended use case(s), is our proposed AI tool appropriate?** This stage would benefit from the expertise of the following stakeholders at the minimum, amongst others: Agency leadership, AI practitioners, frontline workers, community members.

1) Overall goal for using algorithmic tool

Who the tool impacts and serves

- **Who is going to be affected by the decision to use this hypothetical AI tool?**
    - **Who is going to be the most impacted?**
- Who benefits from the use of the tool?
    - To what extent are the targeted outcomes intended to benefit the agency, versus the community?

Intended use

- **What evidence do we have suggesting that the painpoint this tool aims to solve actually exists?**
- **What evidence do we have suggesting that technology may offer a remedy to this painpoint?** (Evidence may include, for example, historical agency metrics, legislature, community members, research reports.)
    - **What evidence suggests the specific form of technology we are envisioning (e.g., predictive analytics) may offer a remedy?**
- What are the additional challenges and risks associated with pursuing a technological solution to this problem?

Involving agency-external stakeholders in determining the goals

- Think about the most impacted stakeholders you identified in response to the questions above. **How do we bring their voices to the table when determining goals?**

- How can we open opportunities for those who are most impacted by the new tool to inform the decision-making process?
- When will we start to engage impacted communities in discussions around how the tool should be designed or used?

Differences in goals

- **Are there differences in the goals the agency versus community members think the tool should address? If so, what are they? If we are uncertain, what can we do to understand potential differences?**
- What evidence do we have that we adequately understand the outcomes the community cares about?
- To what extent are we optimizing the things the agency cares about versus what impacted community members care about?
- Is the process we have in mind for achieving a community-oriented outcome (e.g., child safety) also aligned with the community's desires?

Envisioned harms and benefits

- **(Ideation) What are the potential harms and benefits of the tool, and to whom?**
    - Do benefits outweigh the harms?
    - Do we expect there to be tradeoffs between accuracy, fairness, explainability? For example: making decisions in a completely random fashion may look “fair”, but is not necessarily accurate.
    - Will this tool help us better allocate (scarce) resources?
- **What biases (as a government agency) do we bring into this decision-making process?**
    - How can we identify and mitigate them? What forms of collaboration (e.g., with impacted community members) can help us do this?
- How does this tool help us better deliver to the people we are serving, if at all?

2) Selection of outcomes that the algorithmic tool aims to improve

Impacts of outcome choice

- **Hypothetically, imagine that our tool does a perfect job of improving the outcome that it targets. What additional problems might this create elsewhere in the system?**
- To what extent are we optimizing the things the agency cares about, versus what impacted community members care about?

Assumptions behind outcome choice

- What assumptions are we making, when deciding what the tool should optimize?
- How are we operationalizing goals for the tool, e.g., improving child 'safety'? What assumptions are we making?
- How do we bring providers to the table to decide on the use of outcomes?

Predictability of outcomes

- Have we run any tests on historical data records, to check whether we get predictions on this outcome that actually make sense?
- How rare is the event we are trying to predict? If it is rare, how reliably do we think we can predict it?
- How does the inclusion of additional information (e.g., attributes) improve outcomes?

3) Empirical evaluations of algorithmic tool

Measuring improvement based on outcomes

- **Once the tool is deployed and in use, how can we evaluate how well it is working in the short-term? How can we evaluate how well it is working longer-term?**
- (Ideation stage) What are some ways we might evaluate whether this tool is successful in improving the targeted outcomes?
- (Development stage) For evaluating worker-ADS decisions post-deployment: Do the decisions change by worker experience, worker demographic, or by supervisor?
- (Ideation stage) What performance measures do we plan to use to evaluate the tool?
- (Development stage) What performance measures have already been used in early analyses of historical data, prior to the deployment of the tool?
- Does this tool improve outcomes? How are we operationalizing "improve"?
- How does the use of the tool compare with the status quo? E.g., can we demonstrate the tool improves outcomes for the population of interest?
    - What is the "performance" and "fairness" of the existing baseline/status-quo decision process?
    - Is there someone with relevant domain expertise that could help explain anomalies or trends?
- Do we think there are tradeoffs between accuracy, fairness, explainability? If so, what are they?
- How are we measuring negative and positive impact on families?
- Is there someone with relevant domain expertise that could help explain anomalies or trends?

    - How well do you understand the domain application of the historical data used in evaluation?
    - Are there changes in policies and domain-specific practices in the historical data?
- Are there measured improvements resulting from the model's deployment?
- Are we using appropriate evaluation methods, e.g., synthetic controls, discontinuity analysis when cutoffs on risk exist.
- What outcome measures are we evaluating on? What can these measures tell us, and what can they not tell us?

Centering community needs

- **How can we effectively evaluate the tool from the perspective of impacted community members?**
    - E.g., what does false positive, false negative mean for different impacted communities? How are we weighting false positives and false negatives, in a given use case, based on the relative costs of each type of error for the impacted stakeholders?

Worker perceptions

- **How might front-line workers respond to the tool? How can we better understand their underlying concerns and desires towards the tool?**
- How do front-line workers perceive the algorithm? (e.g., do they consider it a top-down requirement or a useful tool)
- Do domain experts also believe the model 'makes sense', e.g., selection of important features?

Facet 2: Societal and Legal Considerations (54 questions)

The set of questions below are intended to support conversations around the following broader question: **Given the societal, ethical, and legal considerations and envisioned impacts associated with the use of AI tools for our stated goals (identified in Facet 1), is our proposed AI tool appropriate?** This stage would benefit from the expertise of the following stakeholders at the minimum, amongst others: AI practitioners, frontline workers, community members, legal experts.

1) Legal considerations around use of algorithmic tool

- **Do the people impacted by the tool have the power or ability to take legal recourse?**

- Is there clarity around policies, e.g., whether algorithmic outcomes are included under 'public records'
    - If someone asks for information around the tool, but there's no precedent, does the agency know what to do?
- Are you having conversations with the Department of Justice and attorneys, to make sure the new decision models you implement will follow existing policies, procedures, statutes, and rules?
    - Do you know which design decisions will be dictated by the law? For example: In the context of child maltreatment screening, if certain conditions are present in a case, then it is legally required to screen in for investigation.
- Can you inform existing policies, procedures, statutes, and rules to better meet the needs of new decision models?
- Do you need a new temporary rule to receive permission to use the model?
- How are you interpreting challenges to ambiguities in prior legal decisions around the use of the tool?
- What are challenges to interpreting legal documentation and guidelines?
    - How well can we interpret case-specific considerations in the context of legal documentation/guidelines (e.g., when there is a lot of grey in practice, but the law is written in black and white)?
        - E.g., in child maltreatment: "threat of harm" or "physical abuse" allegation type sounds black/white but there are various factors that make this grey. E.g., how hard did it hit them? Did it leave a mark? Action occurred but no impact from the action?

2) Ethical and fairness considerations around use of algorithmic tool

Impacted Community Member Needs

- **Are there differences in the goals the agency versus community members think the tool should address? If so, what are they? If you are uncertain, what are your plans for understanding potential differences?**
    - **What are the envisioned harms and intended benefits from the tool that impact the community and the agency?**
- Can we have impacted community's representatives or advocates at the table, to inform the design and use of the tool?
- How well are we engaging people closest to the problem and those impacted through the entire design, development, implementation, maintenance process?
- Are the outcomes intended for agency or community benefit?
- How well do we understand what outcomes the community wants to improve?

- Do we understand how impacted stakeholders perceive each decision? E.g., emotional valence, potential impacts, etc.
- To what extent are we optimizing the things the agency cares about versus what impacted community members care about?

Involving Impacted Communities

- **What are underlying assumptions that tool developers/researchers may have, regarding the soundness of the design decisions made in the tool?**
- **How can we set up external participation opportunities, to increase access?**
    - E.g., avoiding scheduling during a 9-5pm period (to open involvement to those who want to be involved)
    - E.g., is it possible to involve groups that are not involved and paid by the agency, to get input and feedback?
    - Do we know who should be included? How can we build the right network of people to talk with?
- Who has a seat at the table, to decide how the tool impacts you?
- How are you engaging with people closest to the problem (e.g., frontline workers, community members, or others impacted by the decisions)?
- Have you communicated the limitations and historical context of the data, to community members?
- How well do we understand the costs, risks, and effort required of community members, if we invite them? E.g., many were directly harmed by decisions made by the agency.
- When do we start to engage impacted communities into discussions around the design or use of the tool?

Clarity of Ethics Goals and Definitions

- **Can we agree on a definition of fairness and equity in this context? What would it look like if the desired state is achieved?**

Operationalization of Ethics Goals

- **Are fairness and equity definitions and operationalizations adequately context-specific?** (For example, in the child welfare domain: children with similar profiles receive similar predictions irrespective of race?)
- Do we know how to appropriately operationalize our fairness formulation in the algorithm design?
- Can we mitigate biases in the model?
- How can we balance tradeoffs between false negatives and false positives?

- How well are we integrating domain-specific considerations into the design of the tool?
- Have we recognized and tried to adjust for implicit biases and discrimination inherent in these social systems that might get embedded into the algorithm?

Envisioning Potential Negative Impacts

- **Do we understand the negative impacts of the decision made across sensitive demographic groups?**
- What are the externalities / long-run consequences of the decisions?

3) Social and historical context surrounding the use of algorithmic tool

- **Have we recognized and tried to adjust for implicit biases and discrimination inherent in these social systems that might get embedded into the algorithm?**
- **How might we clearly communicate the limitations and historical context of the data to community members?**
- Are you modeling historical, systemic patterns?

Facet 3: Data and Modeling Constraints (20 questions)

The set of questions below are intended to support conversations around the following broader question: **Given the availability and condition of existing data sources, and our intended modeling approach, is our proposed AI tool appropriate?** This stage would benefit from the expertise of the following stakeholders at the minimum, amongst others: AI practitioners.

1) Understanding data quality

- **How does the data quality and trends compare with an 'ideal' state of the world?**
    - What does our data look like, in terms of different demographic outcomes?
- **Has the definition of the data changed over time?** (E.g., in child welfare, has reunification always meant to reunify with the parent?)
- What data do we have access to?
    - Do we have the data/feature set to replicate the tool/analysis/and predictive accuracy of the existing tool?
- How well do we understand the meaning and value of the data that will be used to train an algorithm?
- How is the quality of this data?
    - How accurate is the data?

    - How recent is the data?
    - How relevant is the data?
    - Has the data been consistently collected?

2) Process of preparing data

- **How are we preprocessing the data?**
- **Who should be involved in making decisions around whether to include or exclude certain data points or features? Do we have plans for involving those people?**
- How do we address bias in the data?
- Do we have metrics for feature importance, that we could show relevant domain experts?
- How well do we understand the data collection process?
    - At what point in time is the data available to you?
    - At what point to use the model?
- Data leakage questions: Are we preventing oversampling of certain populations?
    - E.g., in child welfare: Are we pulling one child per report, and one report per child, to ensure there's no information leakage between training and test sets?

3) Model selection

- Is our model appropriate given the available data? Why or why not?

Facet 4: Organizational Governance Factors (24 questions)

The set of questions below are intended to support conversations around the following broader question: **Given our plans for ensuring longer-term technical maintenance and policy-oriented governance, do we have adequate post-deployment support for our proposed AI tool?** This stage would benefit from the expertise of the following stakeholders at the minimum, amongst others: Agency leaders, AI practitioners, frontline workers.

1) Long-run maintenance of algorithmic tool

Measuring changes in model performance over time

- **Do we expect there will be shifts in performance metrics over time? If so, why? What are our plans for identifying and mitigating those shifts?**
- Do we expect that the data collection process will improve over time? What might this imply for how we maintain the tool? E.g., Is there a need for adjusting thresholds over time?

Mechanisms to identify long-run changes

- Are we repeating feature engineering efforts over time?
    - Are we detecting how trends shift over time at the population level?
- Are there mechanisms in place that track whether certain data features have changed over the years?
- Do we have mechanisms to track longer-term outcomes over time, so that we can monitor for changes in model performance ?
- Do we have the mechanisms to monitor whether the tool is having unintended consequences?

2) Organizational policies and resources around the use of algorithmic tool

Policies around worker interactions

- **Is there training for frontline workers who will be asked to use the tool? What evidence suggests that this training is adequate?**
- How are frontline workers trained?
- Is it clear to workers what information the tool can access, and what information it cannot?
    - How is this communicated to workers?

Governance structures

- **Imagine that we could assemble the "ideal team" to monitor and govern the tool after it is deployed: What are the characteristics of this ideal team?**
    - **Who is the *actual* team that will monitor and govern the tool after it is deployed?**
    - **Given the gaps between the "ideal team" and the actual team we expect to have: What risks to post-deployment monitoring and governance can we anticipate? How might we mitigate these risks?**
- Are there appropriate forms of governance, around the implementation?
    - Do those involved in governance have domain knowledge in the application context and have knowledge of the implementation process?
- Are there sufficient guardrails in place to ensure algorithms wouldn't get weaponized?
    - E.g., IRB-like programs and researchers at the same table, to minimize risk of weaponizing?

3) Internal political considerations around the use of algorithmic tool

- **How well do we understand system administrators' and leadership's perspectives around the use of this tool?**
- **How well do staff and leadership understand 'why' the tool could bring value?**
- (Ideation phase) Do system administrators and leadership perceive this tool positively?
- (Ideation phase) Do leadership support the future use of the tool?
    - Do we have backing at a leadership level? E.g., director, agency, governor, community partners?
- (Ideation phase) Is there sufficient buy-in from middle managers and executive support?
- Do we have mechanisms to address concerns that could come up during the ideation and design process?

Scratchpad: External resources to potentially link towards from the guidebook

For documenting information about the model and dataset

- A People's Guide To Tech - Allied Media Projects
- Datasheets for Datasets
- Model Cards