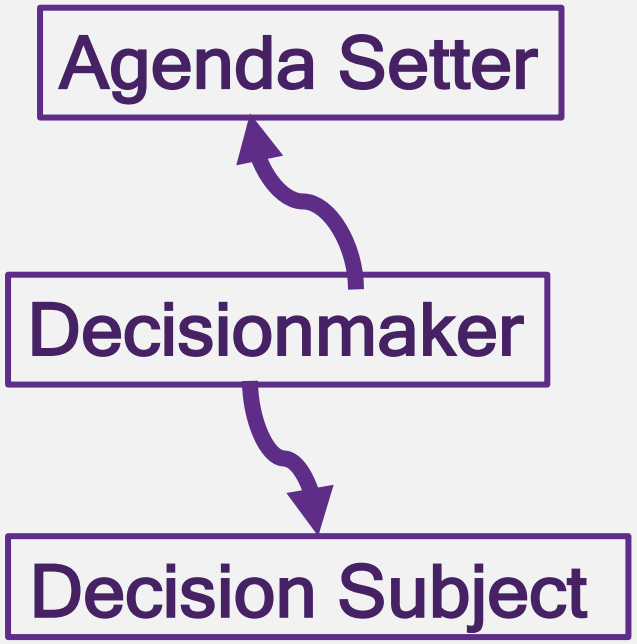
Explanation as Justification: Traditional View

- Human Decisionmaking Processes are Black Boxes/Inscrutable
  - Decisionmakers' mental processes are hidden and unobservable
  - Humans cannot/ may not wish to report them accurately
  - Reason-giving ("explanation") is required precisely because of this "black box" problem

- Explanation is a mechanism for justifying decisions
  - Not simply (or necessarily) describing the actual decision process
  - Explaining how the decision comports with acceptable rules/standards
    - May be substantive or procedural

- Explanations contribute to decision/decision system
  - Legitimacy
  - Accuracy/appropriateness
  - Evaluation, review and correction
    - of case-by-case decisions
    - of rules/standards

(Overly) Simple View of Explanation in Decision System



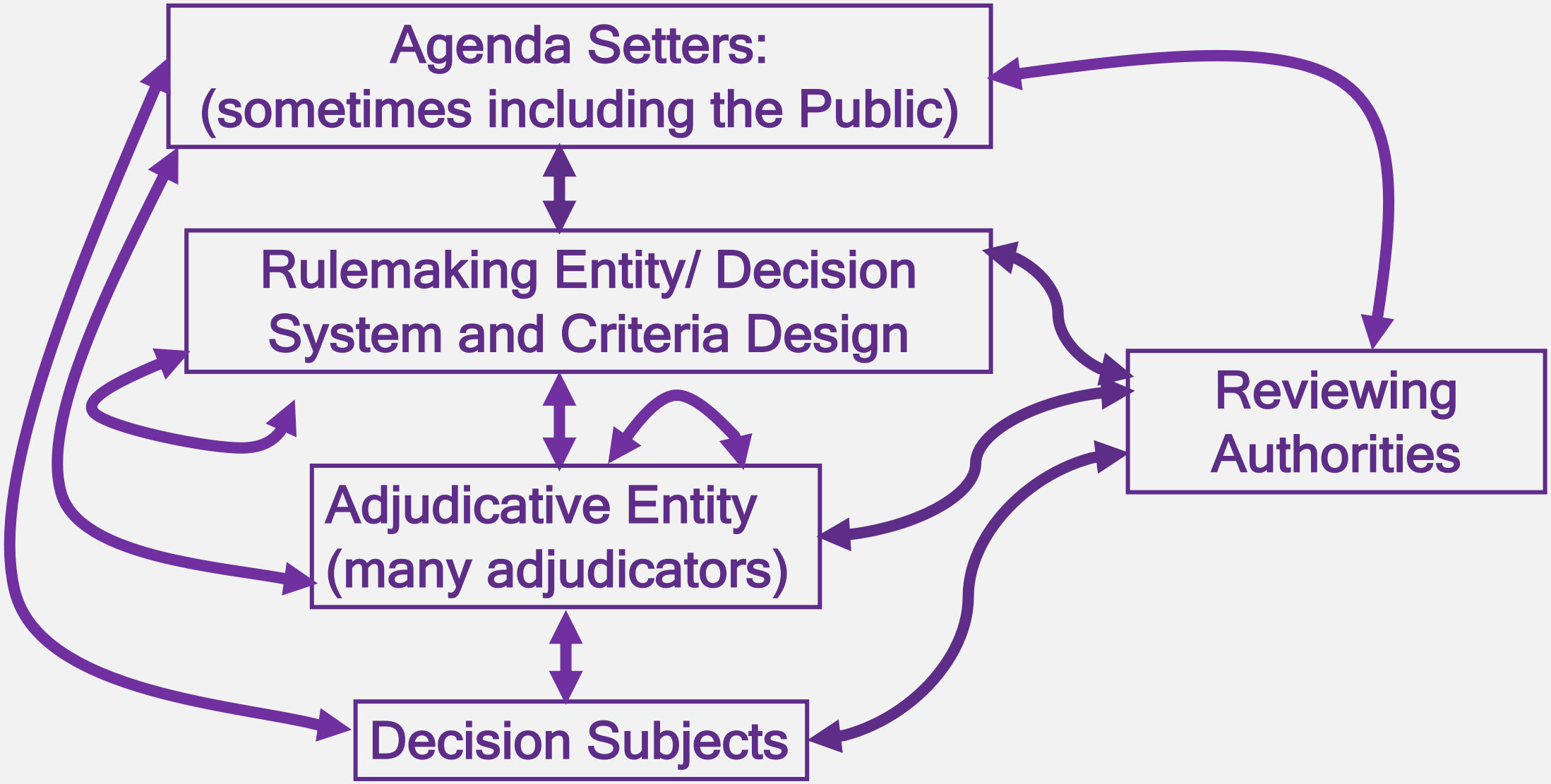
- Explanation as justification
  - Even when the “black box” is a human
- Explanation serves functions such as:
  - Legitimacy
  - Guiding decision subject behavior

Explanation in a Delegated and Distributed Decision System

When decisions must be made in many comparable cases:

- Decisionmaking is usually delegated
  - Purpose and goals set by agenda-setter
  - Rulemakers create rules/standards to be applied by
  - Adjudicators to case-specific facts
- Decisionmaking is usually distributed
  - Between rulemakers and adjudicators
  - Among (many) adjudicators
- Delegation and distribution create:
  - Principle-agent problems that require accountability
  - Coordination problems
  - Both of these problems are traditionally addressed (at least in part) by explanations between various decision system actors

Explanation in a Delegated and Distributed Decision System: Many Flows and Purposes



Explanations in Delegated, Distributed Decision Systems

Functions of explanations in such systems include:

- Legitimacy to decision subjects
- Guidance to decision subjects
- Ensuring that rulemakers and adjudicators are accountable to agenda setters:
  - No misunderstandings/mistakes
  - No avoiding effort
  - No conflicts of interest
- Coordination within rulemaking entity
- Coordination between rulemakers and adjudicators
- Coordination among and consistency between adjudicators
- Allowing for correction of adjudicative outcomes by reviewing entities
- Facilitating updating and improvement of rules
- …

Automated Decision Tools: Explanation Issues for Rulemakers

- Specify which decision criteria should be automated
  - What outcome variables to use? What features to use?
  - Is the available data adequate to the task for all decision subjects and sub-groups of decision subjects?
  - Are there appropriate metrics and benchmarks for validation (and continued evaluation) of tool performance?
  - Requires communication between data scientists and domain experts
- Specify criteria for human adjudicators to evaluate
- Specify how human adjudicators are to use automated tool output
  - What do they need to know in order to use the tool output appropriately?
- How can adjudicators provide feedback to rulemakers about how the automated tool is working in real cases in an evolving world?
- How can agenda setters evaluate the decision system?

Automated Decision Tools: Explanation Issues for Adjudicators

- Adjudicators need adequate explanation of automated tool output
- Adjudicators may need to explain their decisions to:
  - Decision subjects
    - Legitimacy
    - Guiding future behavior
  - Agenda setters
    - Accountability, Rule evaluation
  - Reviewing bodies
  - Rulemakers and other adjudicators
    - A means of vetting how the rules apply in practice
    - Catching generalizability and bias problems in application
    - Checking for meaningful decision consistency

- Do these diverse explanatory flows require diverse technical approaches?
- Explanations are not (necessarily) descriptions: they are mechanisms for justifying decisions
    - Can other forms of validation/verification (sometimes) serve the same functions?
- When might the need for these other explanatory flows mean that automated decision tools should not be employed?