Decisionmaking, Machine Learning and the Value of Explanation

Katherine Strandburg
New York University School of Law

The Requirement to Explain Decisions*

- **Procedural due process**:
  - Individuals subject to government decisionmaking are entitled to appropriate procedural protections
  - Required protections vary and the level of procedure required depends on:
    - (1) the private interest that will be affected by the official action
    - (2) the risk of an erroneous deprivation of such interest through the procedures used, and probable value, if any, of additional procedural safeguards; and
    - (3) the Government's interest, including the fiscal and administrative burdens that the additional or substitute procedures would entail.

* Warning and apology: My legal references are quite US-centric. But the underlying principles are general

The Requirement to Explain Decisions

- **Explanation is a core aspect of due process:**
  - Judges generally provide either written or oral explanations of their decisions
  - Administrative rulemaking requires that agencies respond to comments on proposed rules
  - Agency adjudicators must provide reasons for their decision to facilitate judicial review
  - ….
- **When explanation is not required:**
  - Jury decisions – made by "peers"
  - Legislative enactments – democratic legitimacy
  - Government actions without significant impact or with good reasons not to explain (i.e. investigations)

Two Sorts of Explanations

- **Descriptive explanation:**
  - How did decisionmaker X arrive at outcome Y?
  - Descriptive, not normative
  - Potential critiques:
    - Based on incorrect empirical facts
    - Logical mistakes in legal analysis
    - Not credible
- **Justification:**
  - Why is outcome Y the right decision?
  - Normative
  - Potential critiques:
    - Disagreement about approriate normative values
    - Not persuasive

Aspects of Legal Decisionmaking

- **Legal interpretation:**
  - Almost never entirely straightforward
  - Usually has normative aspects
  - Requires both
    - Descriptive explanation
    - Justification
- **Applying Law to Particular Facts:**
  - Two steps:
    - Fact-finding
    - Using a given legal interpretation in conjunction with the facts to derive a decision
  - Requires only descriptive explanation

Why Require Explanations?

- Improve Decisionmaking Accuracy
- Promote Fair and Unbiased Decisionmaking
- Promote Legitimacy and Trust in Social Institutions
- Promote Compliance with Law
- Respect Individual Dignity and Autonomy

Improving Decision Accuracy

- What does "accuracy" mean?
  - Correct legal interpretation
    - Consistent with text of the rule or statute
    - Appropriate method for explicating remaining ambiguities
    - Uses appropriate normative considerations where necessary
    - Is analytically sound
  - Correct application
    - Relies on accurate and relevant empirical facts
    - Uses correct legal interpretation
    - Is analytically sound

Improving Decision Accuracy

- How can explanation improve accuracy?
  - The exercise of explaining helps decisionmakers to catch and avoid errors
  - Making explanations available to others incentives careful decisionmaking
  - Explanations provide a basis for disputing decisions and for review by higher authorities
  - Explanations, especially cumulatively, promote robust legal development by
    - facilitating critique and debate
    - Highlighting situations in which current legal interpretations or rules lead to problematic outcomes
- Both descriptive explanations and justifications can improve accuracy for these reasons

Promoting Fair and Unbiased Decisions

- **Unfair or biased decisions stem from:**
  - Pernicious explicit motivations
  - Implicit or unconscious bias
  - Unanticipated results of applying legal interpretations
- **Pernicious explicit motivations**
  - Decisionmakers will lie about their reasons
  - Attempts to obfuscate true motivations may result in less persuasive or analytically sound explanations
  - Decisionmakers who recognize this may be deterred from acting on illicit motives
  - If they are not deterred, their implausibility of their explanations may lead reviewers to overturn their decisions
  - Of course, this won't always work

Promoting Fair and Unbiased Decisions

- **Implicit bias**
  - May also lead to unconvincing explanations
  - Decisionmakers may recognize this for themselves and modify their decisions
  - Reviewing authorities are more likely to revers
  - Also not guaranteed to work
- **Unintended consequences of correct application of legal rules**
  - Explanations, cumulatively, may highlight biased or unfair outcomes, promoting reform
  - Also may not work

Promoting Legitimacy and Social Trust

- **Empirical studies show that "procedural justice" promotes more favorable views of decisionmaking processes**
  - **Explanations are an aspect of procedural justice that are likely to have this effect**
  - **Procedural justice has an evil twin: complaceny in the face of substantive injustice!**
  - **E.g. Provide an elaborate hearing, listen to an individual's arguments, then make an unjust decision**
  - **Explanation-giving is hard for an evil twin**

Promoting Legal Compliance

- **Explanation clarifies legal requirements and makes it easier to comply**
  - For the subject of the decision who will face similar situations in the future
  - Cumulatively, for everyone, especially when explanations are aggregated by some intermediary
  - Of course, this assumes that promoting legal compliance is a good thing!
- **Is gaming the system compliance's evil twin?**
  - Rule of law: citizens ordinarily have the right to know the law and comply strictly with the letter of the law
  - Gaming the system is only possible for decisions made on discretionary grounds, where compliance is not the goal (e.g. targeting investigations)

Promoting Dignity and Autonomy

- Explanations of decisions are inherently valuable because they show respect for the dignity of those affected
- Explanations enhance autonomy by giving individuals options about whether and how to comply with the law
- Explanations enhance dignity by treating individuals as democratic citizens rather than subjects

Explanation and Automated Decisionmaking

- Are there substitutes for explanation in the context of automated decisionmaking?
- Do explanations serve the same purposes for automated decisionmaking?

Improving Decision Accuracy

- **Automation improves accuracy in one particular respect without relying on explanation**
  - Given a well-defined legal interpretation and a well-defined set of "facts" (data), automation ensured that legal application is analytically sound
- **But may diminish accuracy in other respects**
  - Legal interpretations must be put into codable form and communicated to programmers
    - This warp the process of legal interpretation and obscure normative considerations
  - Legally relevant factual situations must be represented in terms of available data proxies
  - Without explanations, cumulative outcomes may not facilitate reform

Promoting Fair and Unbiased Decisions

- **Pernicious explicit motivations and implicit bias**
  - Computers do not have pernicious motivations or implicit biases
  - But pernicious motivations and implicit biases can affect the human activities of encoding legal interpretations and selecting factual data
  - Automated decisionmaking offers some opportunities to encode metrics for fairness and bias into the system, which can be used to evaluate and improve decisionming
  - The selection of such metrics is a normative value judgment, involving tradeoffs between these and other values
  - Such selections should be justified by explanations

Promoting Fair and Unbiased Decisions

- **Unintended consequences of correct application of legal rules**
  - **Without either explanations or some other form of ex post analysis, automated decisionmaking processes will not detect such cumulative unintended consequences**

Promoting Legal Compliance

- **Some ways of encoding a legal rule require precise specification**
  - **If such encoded rules are disclosed, they can promote compliance with the encoded interpretation of the rule**
  - **The bottom line depends on the validity of the encoded interpretation**
- **Rules resulting from machine learning may not be interpretable or may have interpretations that are not easily translated into behavior**
  - **In such cases, automated decisionmaking does not promote legal compliance**

Promoting Legitimacy and Social Trust

- Kroll et al suggest computation methods to certify that automated decisionmaking has followed a prescribed automated
  - Such accountability will enhance legitimacy and trust
- These methods do not ensure appropriate legal interpretation or accurate factual data
  - Without explanation, legitimacy and trust may decrease
- Transparency alone is not justification
- Statistical correlation may not provide sufficient justification to promote legitimacy and trust

Promoting Dignity and Autonomy

- Explanations play the same part in promoting dignity and autonomy for automated decisions as they do for traditional decisionmaking
- Some versions of interpretability will not provide the kinds of justifications needed for these purpose