Accountability, Deterrence, and Identifiability

Aaron D. Jaggard
U.S. Naval Research Laboratory

Collaborators

Joint with Joan Feigenbaum, Rebecca Wright

- Parts also with Jim Hendler, Danny Weitzner, and Hongda Xiao

Overview

- “Accountability” is used in lots of ways
  - Complements preventive security
  - Is generally viewed as good
  - This space needs formal definitions
- What are the goals?
- Formalize accountability and related notions
  - Facilitate reasoning about this
  - Enable comparison (and discussion) across sytems
- What is the role of identity/identifiability?

What Are the End Goals?

- Deterrence
  - Focus here
- Others
  - Comply with requirements on processes
- With an eye on these, don’t make things implicit that don’t need to be

Temporal Spectrum [FJWX'12]

- Many systems focus on various (different) aspects of evidence/judgment/punishment
  - E.g., accountable blacklistable credentials, accountable signatures, e-cash, reputation schemes, ...

Focus on Punishment

- Shift focus
  - At least if end goal is deterrence
  - *Effect on* violator instead of *information about*

- Reduce the built-in identity assumptions
  - May still want to use evidence, etc., but don't want this (and especially identity) implicit in definitions

Comments on Accountability

- “Accountability is a protean concept, a placeholder for multiple contemporary anxieties.” [Mashaw]
- “[A]ccountability has not yet had time to accumulate a substantial tradition of academic analysis. ... [T]here has been little agreement, or even common ground of disagreement, over the general nature of accountability or its various mechanisms.” [Mulgan]

A CS Definition of Accountability

- "Accountability is the ability to hold an entity, such as a person or organization, responsible for its actions." [Lampson]

Working Definition [FJW'11]

An entity is *accountable* with respect to some policy (or *accountable for* obeying the policy) if, whenever the entity violates the policy, then, with some non-zero probability, it is, or could be, punished.

Working Definition of Accountability

- Builds on definition of Lampson
  - Avoids “hold ... responsible”---explicitly don’t require external action
- Shift focus from evidence and judgment to punishment
  - Reduce need for identity
  - May want to reserve “accountable” for when violator is identified [Weitzner]
    - Need to be able to distinguish those cases

Punishment Desiderata

- Unrelated things events shouldn't affect whether a violator is viewed as being punished
  - "Luck" shouldn't affect things
  - Punishment should be related to violation in question

Automatic and Mediated Punishment

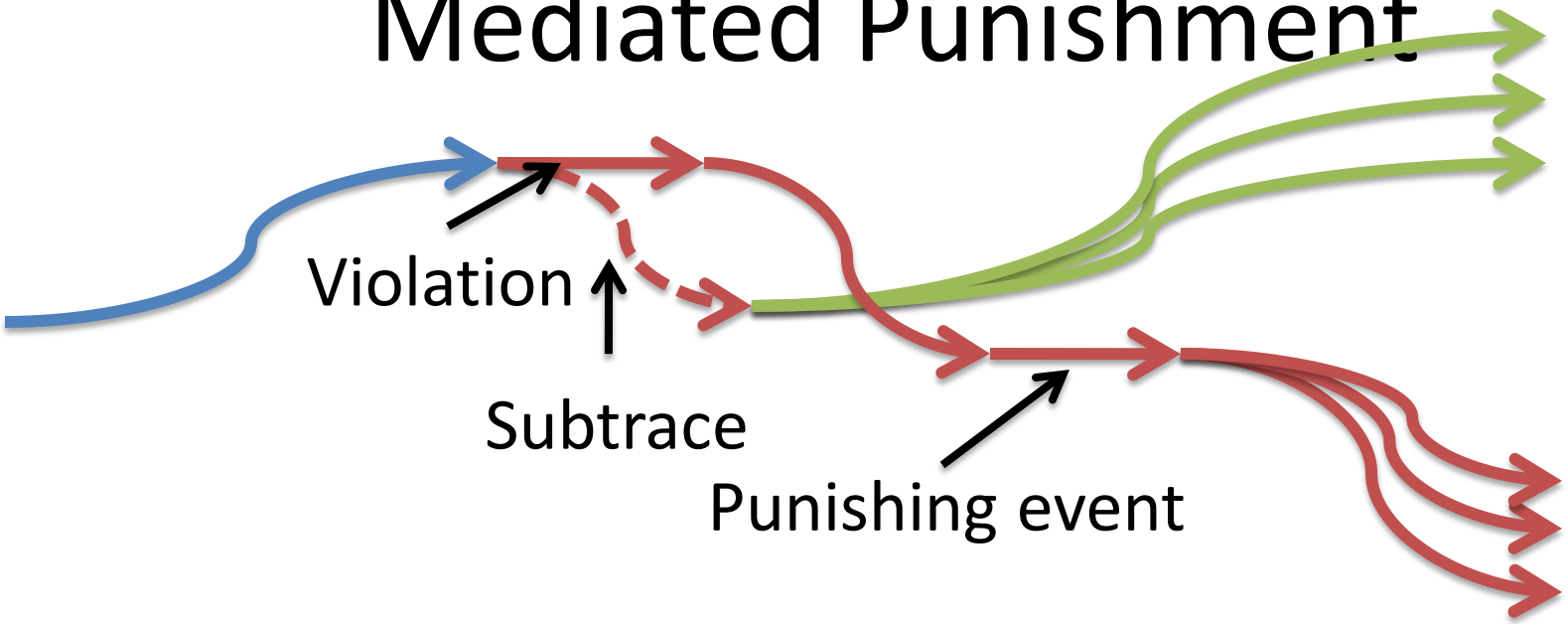
Intuitively:

- Punishment should be connected to the violation
- Punishment could be mediated by the action(s) of some authority responding to the violation (or a conviction)
- Punishment could happen without any punishing act being done
  - This might reduce need for identifiability!

Formal Model (Outline) [FJW'11]

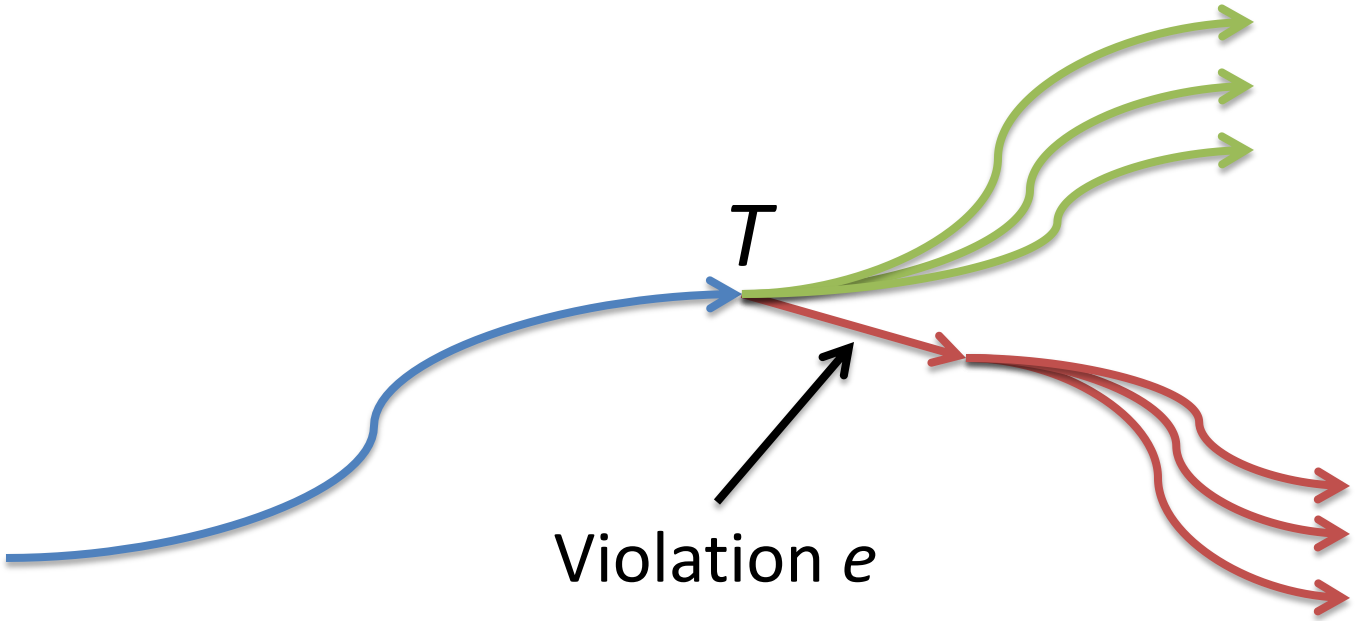
- System behavior as event traces
- Utility functions for participants
  - Maybe only know distribution or "typical" utility
- Principal(s) associated to events
- What qualifies as "punishment" for a violation?

Mediated Punishment



- Punishing event must be caused by the fact of the violation
- Compare outcomes after punishment with those without the punishment
  - Need to remove other events caused by the violation

Approach: Automatic Punishment



- The violation is automatically punished if the violator's utility is lower in the outcomes extending the violating trace *Te* than in the outcomes extending *T* but not *Te*.

Example: Three Strikes

- Using “typical” utilities, we capture the idea that even sociopaths are punished
  - Expected utilities might be skewed
- What is “effective” punishment?

Open/Closed Systems

- Degree to which a system is “open” or “closed” appears to be important
- Principals, identities/nyms, and systems
  - What does system boundary require of mapping between principals and identities
    - Computational or algebraic restrictions
- Example potential tradeoff
  - Deterrence may be effective if we punish the nym
    - Contexts where being able to act *using that identity* is important
    - Wouldn’t need to be able to invert mapping
  - Other times, may need more (computable) information about mapping

Primitives

- Temporal spectrum (evidence, judgment, punishment)
  - What do these need to provide abstractly? E.g., to what extent must evidence be independently verifiable?
- Blame/violation
- Causality
- Identity-related
  - Binding between principals and identities; linkages between actions

Other Dimensions [FJWX'12]

- Information
  - Is identity required to participate in the system?
  - Are violations disclosed? How broadly?
  - Is the violator identified? How broadly?

- Action
  - Centralized v. decentralized (generally and in response to violations)
  - Automatic v. mediated
  - Requires continued access to violator?

Social Aspects

- Should we reserve “accountability” for approaches that require identification? That might be consistent with common uses of “to hold someone accountable.”
  - This may not be the fundamental goal; we may really be after deterrence.
  - One can be deterred even if one will not be identified
- Possible approach: Allay user concerns by promoting “deterrence” instead of “accountability”

Questions

- Are these the right notions to formalize?
  - Will this framework be useful for proving things?
  - Capturing identity?
- What about related notions
  - Compensation, detection/diagnostics, authorization
- Open/closed systems
  - Compare with international-relations example of non-accountability
- Subset/delegated accountability
  - Don't (immediately) have individual punishment
  - Reduce level of identifiability
  - How to induce participation?