Towards a Formal Model of Accountability

Joan Feigenbaum*
Department of Computer Science
Yale University
New Haven, CT USA
joan.feigenbaum@yale.edu

Aaron D. Jaggard†
Department of Computer Science
Colgate University
Hamilton, NY USA
adj@dimacs.rutgers.edu

Rebecca N. Wright‡
Department of Computer Science and DIMACS
Rutgers University
New Brunswick, NJ USA
rebecca.wright@rutgers.edu

ABSTRACT

We propose a focus on accountability as a mechanism for ensuring security in information systems. To that end, we present a formal definition of *accountability* in information systems. Our definition is more general and potentially more widely applicable than the accountability notions that have previously appeared in the security literature. In particular, we treat in a unified manner scenarios in which accountability is enforced automatically and those in which enforcement must be mediated by an authority; similarly, our formalism includes scenarios in which the parties who are held accountable can remain anonymous and those in which they must be identified by the authorities to whom they are accountable. Essential elements of our formalism include *event traces* and *utility functions* and the use of these to define punishment and related notions.

Categories and Subject Descriptors

K.6.5 [**Computing Milieux**]: MANAGEMENT OF COMPUTING AND INFORMATION SYSTEMS—*Security and Protection*

General Terms

Security, Theory

Keywords

Accountability

*Supported in part by NSF grant 1016875 and DARPA contract N66001-11-C-4018.

†Also affiliated with DIMACS, Rutgers University. Supported in part by NSF grant 1018557.

‡Supported in part by NSF grant 1018557.

NSPW'11, September 12–15, 2011, Marin County, CA, USA.

1. INTRODUCTION

The Computer Science community's dominant approach to information security has typically been *preventive*: Before someone can access confidential data, connect to a private network, or take any other security-sensitive action, she should be required to prove that she is authorized to do so. As the scale and complexity of online activity has grown, it has become increasingly apparent that the preventive approach by itself is inadequate. In response, several researchers, including Lampson [14] and Weitzner *et al.* [23], have suggested that, in addition to maintaining its traditional goal of preventing breaches of security policy, the community should embrace the complementary goal of *accountability*: When an action occurs, it should be possible to determine (perhaps after the fact) whether a policy has been violated and, if so, to punish the violators in some fashion. To see why a purely preventive approach to information security in inadequate, consider the following three archetypal scenarios:

Copyright: Digital distribution of copyright works provides a clear example of the need for a more flexible toolkit. In an attempt to prevent piracy—*i.e.*, large-scale illegal copying or redistribution—distributors use digital-rights-management (DRM) systems. These systems generally subject all users to limits (or even prohibitions) on copying and modification of the works in question, regardless of whether the actions that a user wants to take would be considered fair use under copyright law. Some fair-use advocates attribute this strategy purely to greed on the part of rights holders, but there is a technical reason for it as well: Because there is no way to distinguish *a priori* between a legitimate user and a pirate, the only way that a rights holder can prevent piracy is to impose technical limitations on *all* copies of the work that he distributes. In arguing for the preservation of fair use in digital-distribution regimes, some advocate that these universal technical limitations should not be too strict, and some advocate for the right to "circumvent" such limitations in order to make use of the material in a manner that complies with copyright law even if it violates the policies of the DRM system. Neither of these approaches is entirely satisfactory: On the one hand, there is no reason to believe that even the most generous of usage terms will prevent large-scale, illegal copying of valuable entertainment content; on the other hand, it is unfair to expect people who want to make fair use of copyright material but are not skilled programmers to hack around DRM systems or even to know how to obtain and use circumvention tools.

The essence of the problem is the preventive approach to enforcement of copyright law. While it may indeed be impossible to distinguish *a priori* between a legitimate user and a pirate, copyright law could still be enforced if one could hold users accountable *a posteriori* for the uses that they make of copyright works. Note that this approach worked quite well for many years in the print world; for example, physical books have always been sold without DRM systems, but book publishers can sue people who engage in large-scale illegal copying and distribution of books, because large-scale activity of this sort is detectable (*e.g.*, in a money trail and occasionally in broad daylight on urban sidewalks). In the digital world, publishers have tried without much success to use watermarking systems to track pirates. Are there other protocols for accountability in the digital-copyright world that would be more successful?

Surveillance: The desire to stop illegal wiretapping and surveillance faces obstacles similar to those faced in the copyright scenario. Law-enforcement and intelligence agencies need warrants in order to eavesdrop on US citizens[1] but not on certain other people. Of course, it is infeasible even to determine the endpoints of many Internet traffic streams, much less to determine whether the sender and recipient are US citizens. This dilemma has led some people to conclude that eavesdropping (on Internet communications, anyway) is inconsistent with fourth-amendment protection against unreasonable searches and that, therefore, either the fourth amendment or government eavesdropping on Internet communications must be abandoned altogether. Neither would need to be abandoned if there were a way to grant law-enforcement or intelligence agents temporary access for the purpose of determining whether the sender and recipient are US citizens; an agent who determined that a warrant is needed would be required to go to court to get one and not to use the traffic he eavesdropped on temporarily for any other purpose. Although it is not immediately clear how to hold agents accountable (in a way that is practical and effective) for following these requirements, it may nonetheless be possible to do so.

Web search: Search plays a central role in a user's Internet activity, and thus users reveal a great deal of personal information to search engines. Google, in particular, is well aware that some people are disturbed by that fact and has striven mightily (and so far largely successfully) to convince them that it handles this information properly. Nonetheless, it would be highly desirable if search companies could be held accountable for their uses of personal information. Although this is not a purely technical problem, because there is no widespread social agreement about what the legitimate uses of search data are, it is clear that a purely preventive approach will not work to "secure" personal information that users reveal when searching the Web. Search technology and the services offered by search companies are evolving rapidly, and users should not have to opt out whenever they have (justified!) concerns about potential misuse of their data; rather, accountability technology should be developed and deployed along with search technology, social networking, and other services that depend on personal data.

[1]Warrants are actually needed in order to eavesdrop on "US persons," which is a broader group than US citizens; because the definition of "US person" is complicated and not necessary to make our point about accountability, we omit it from this discussion.

We propose a security paradigm in which accountability is considered along with a preventive approach to security. Although there is very widespread agreement that "accountability" is an important aspect of information security, there is no agreement on a precise definition of the term, and indeed different researchers use it to mean different things. For example, several influential experimental works, *e.g.*, [1, 16–18], require that system participants have persistent identities that are known to those who hold them accountable; this precludes the possibility of an accountable information system in which participants are anonymous. At the opposite extreme, accountability is sometimes formulated *only* for scenarios in which honest participants remain anonymous, *e.g.*, [3–6, 21]; in these works, one is held accountable precisely in the sense that one's identity can be exposed if one violates the prescribed security policy or protocol. Neither of these approaches is sufficiently general for the plethora of online interactions in which a robust notion of accountability is desirable.

In this paper, we present a new, formal model of accountability based on *event traces* and *utility functions*. Intuitively, participants are held accountable in the sense that they derive lower utility from traces that include security-policy violations for which they are responsible than they do from those in which all of their actions are policy-compliant. Our contributions include the following:

- We provide a formal definition that is more general and potentially more widely applicable than the accountability notions that have previously appeared in the security literature, *e.g.*, those in [3–6, 21]. In particular, we treat in a unified manner scenarios in which accountability is enforced automatically and those in which enforcement must be mediated by an authority; similarly, our formalism includes scenarios in which the parties who are held accountable can remain anonymous and those in which they must be identified by the authorities to whom they are accountable.

- We handle both scenarios in which there is a natural distribution on traces (and thus a natural notion of expected utility) and those in which there is no such distribution. In the latter cases, we use ranking functions (as used by Halpern [9] in defining causality) to identify typical utility.

- We illustrate and explain the differences and similarities between "accountability" based on utility functions and the economic notion of "incentive compatibility." In particular, a potential policy violator may have a utility function under which he benefits (in expectation) from violating the policy and another utility function under which he does not benefit in expectation. Depending on the which of these is more typical or, in a probabilistic model, what the distribution of these is across potential violators, we may still be able to say that violators are held accountable for the violating the policy in question.

We start in Section 2, with a brief review of some of the prior work on accountability and a high-level explanation of

one of the main ingredients of our approach—*i.e.*, a flexible notion of what it means to "punish" an entity that is accountable for obeying a security policy but in fact violates that policy. In Section 3, we formalize the notion of punishment and then use it to develop our accountability model, handling both automatic and mediated punishment and both expected and typical utility; we also compare and contrast accountability with incentive compatibility. In Section 4, we examine further the connection between punishment and accountability and give examples of scenarios that are readily handled in our model but not in prior accountability models. We conclude with additional discussion in Section 5.

2. DEFINING ACCOUNTABILITY

The concept of accountability is central to many activities and arrangements in government and business, including, for example, elections, work-place hierarchies, delegation of authority, and fiduciary responsibility. As a result, accountability has been studied in law, political theory, and philosophy. More recently, there has been attention to accountability from computer scientists.

2.1 Related work: Social sciences

A comprehensive review of this literature is beyond the scope of this paper, but we highlight several points that are particularly relevant.

Grant and Keohane [8] study accountability in the interaction of nation states; they define it as the "right of some actors to hold other actors to a set of standards, to judge whether they have fulfilled their responsibilities in light of these standards, and to impose sanctions if they determine that these responsibilities have not been met." They point out that their approach presupposes an international framework within which nation states interact; one nation's unilateral defense of its own interests is not viewed as an accountability mechanism. We believe that a technological realization of Grant and Keohane's definition could be quite useful but that it is not sufficient for all online interactions that require accountability. For example, it tacitly assumes that all of the actors have persistent identities and are known to each other.

Broad-ranging social and legal theories of accountability often lack precise definitions. Indeed, the absence of precise definition has been noted by multiple authors. Mashaw [15], who examined the concept in administrative law, states that "[a]ccountability is a protean concept, a placeholder for multiple contemporary anxieties." Mulgan [19], whose domain is democratic, national governments, notes that "accountability 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." In an early, prescient study of "computerized society," Nissenbaum [20] draws on philosophical analyses of moral blame and responsibility to identify barriers to accountability in software development and deployment.

2.2 Related Work: Computer Science

Lampson [13] put forth a definition that is similar to that of Grant and Keohane [8]:

> Accountability is the ability to hold an entity, such as a person or organization, responsible for its actions.

We take this definition, which is useful but neither sufficiently precise nor sufficiently general, as our point of departure in Sec. 2.3 below.

As noted in Sec. 1, there is a considerable amount of prior work in computer science in which "accountability" is achieved by devising protocols in which actors who violate the security policy may have their identities exposed, while actors who comply with the policy are guaranteed to remain anonymous. These include, for example, [3–6, 21].

Similarly, there is a great deal of prior work that relies upon actors' having persistent identities that are known to all concerned. In addition to the experimental work mentioned in Sec. 1, there has been some theoretical work along these lines. For example, Küsters *et al.* [12] give an approach in which systems deliver verdicts consisting of Boolean formulæ built up from assertions that identified agents are dishonest. Bella and Paulson [2] take an approach in which the goal is to produce long-lived evidence, usually digitally signed, of the actions of a party who is to be held accountable by his peers. Jagadeesan *et al.* [11] take an approach in which auditors are able to "blame" senders of messages that deviate from the prescribed protocol.

In all of these examples, the system provides a mechanism for identifying those who have misbehaved, but leaves it external to the system to determine whether and how to actually hold the misbehavers accountable. In contrast, we seek a more flexible definition that does not necessarily require identification to happen, and in which the accountability mechanism might be part of the system itself, rather than relying on a judge or external, out-of-band third party to enforce the accountability.

2.3 Towards a formal definition of accountability

Lampson's definition provides a useful starting point; however, to the extent that "hold[ing] an entity ... responsible for its actions" suggests a restriction to some sort of *active* enforcement, we want a broader definition. We start with the following "working definition" of accountability.

Working Definition 1. (Accountable entity) 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.

In particular, we want to explicitly allow for the possibility that there is not another entity that "hold[s the violator] ... responsible for its actions." This aspect of our approach is especially important if we wish to explore how accountability might be achieved without the level of identifiability that is typically assumed to be required for accountability.

With the goal of formalizing such a definition, we first formalize the notion of punishment, in Sec. 3. As we discuss, "punishment" is itself open to multiple interpretations; we discuss approaches to this below. We do not require that punishment actually be inflicted. For example, a police department might decide not to pursue minor crimes it discovers during a long-term investigation in order not to compromise its surveillance of individuals suspected of major crimes. The people who commit those minor crimes are still accountable in the sense of this working definition—they could be punished, even though, in this case, a conscious decision is made not to punish them. Our framework also in-

cludes both systems in which violators are always punished and those in which they are punished sometimes but not always (which might be sufficient to deter violations). Our working definition allows, but does not require, the punishment to happen automatically, without any active involvement by other entities.

3. FORMALIZING PUNISHMENT

We stated our working definition of "accountability" (Working Definition 1) in terms of "punishment." While we have intuitive ideas about what does and does not constitute punishment, we need a formal approach to provide a foundation for a discussion of accountability. We start by describing the framework that we will use to formalize punishment and related concepts and then carry out that formalization.

At an intuitive level, we might use the following English-language rule-of-thumb to describe punishment.

> Punishment means that the violator is worse off—in either an expected or a typical sense—after the punishment than he would have been had he not committed the violation.

This may be specialized in multiple ways to formalize what is meant by "worse off." As described below, we frame this in terms of decreased utility; that, of course, gives rise to multiple questions about how to measure the decrease in utility, whether punishment is truly effective, *etc.* We focus on definitions using either probability or typicality, also discussed below; even a small set of options leads to many reasonable (and useful) definitions of punishment (and, in turn, accountability).

In defining punishment, we wish to satisfy multiple goals:

- "Bad luck" should not qualify as punishment. The utility of a principal might be decreased for reasons—such as a side effect of some other actions in the system—that are unrelated to a violation committed by the principal. Such cases should not be classified as punishing the principal in question.
- Similarly, "good luck" should not need to be completely undone in order to punish a violator. Someone might violate a policy and then receive an unexpected windfall that is completely unrelated to the violation. In punishment, the violator's utility should be decreased relative to its value after the windfall (but without effects of the violation) instead of relative to the value of the utility without the windfall.
- Punishment intended to be of a particular policy violation should not also qualify as punishment of other actions by that principal. For example, saying that a violation is punished whenever the violator's expected utility is subsequently decreased would mean that punishment of a violation also punishes all actions done by that principal before the time of the violation.

With these informal desiderata in mind, we now turn to our formal approach to punishment. Sections 3.1 and 3.2 provide the basic formal definitions that we later use to define punishment. In Sec. 3.3 we define a notion of "automatic punishment" in multiple ways (depending on whether traces and utilities are viewed probabilistically or typically); as an application of this, we highlight the second-price auction, which illustrates that our definition of punishment can be satisfied without identifying the violator (or even knowing that a violation has occurred). This suggests that identity and accountability may not be as tightly coupled as is often thought. We also discuss distinctions between this notion of punishment and the economic notion of "incentive compatibility." We then define, in Sec. 3.4 a notion of "mediated punishment," which includes the (perhaps more intuitive) concept of a punishing action. However, the formalization of this is more subtle than the formalization of automatic punishment.

3.1 Traces and utilities

We start by defining traces as sequences of events; those that cannot be further extended are the outcomes of a system. The benefit (or utility) derived by different participants is defined, as usual, on outcomes. We will be interested in the way that different events (such as attempts at punishment) affect the utility of participants (such as those who have violated a policy). As a result, we are also interested in the extension of utilities to all traces and not just outcomes; we discuss two approaches to this, one probabilistic and one in terms of "typical" outcomes extending a trace. These allow us to compare the utility (either expected or typical) that a participant would receive before and after a potentially punishing action. It is useful to have both approaches because we may not know the distribution on utilities or traces, and we may wish to ignore extreme cases.

Definition 1. (Events, traces, and enabled events) There is a collection $\mathcal{E}$ whose elements are called *events*. There is a set $\mathcal{T}$ whose elements are finite sequences of events (but not necessarily all such sequences); the elements of this set are called *traces*. We require that every prefix of a trace is also a trace; so, if $\mathsf{a}, \mathsf{b}, \mathsf{c} \in \mathcal{E}$ and $\mathsf{acb} \in \mathcal{T}$, then $\mathsf{a} \in \mathcal{T}$ and $\mathsf{ac} \in \mathcal{T}$ as well. We say that an event e is *enabled* at a trace T if Te (*i.e.*, T followed by e) is also a trace; continuing the previous example, if $\mathsf{acbe} \in \mathcal{T}$ but $\mathsf{acbd} \notin \mathcal{T}$, then e is enabled at $\mathsf{T} = \mathsf{acb} \in \mathcal{T}$, but d is not enabled at T.

We think of a trace as capturing the events that occurred in a system. In our current formalization, traces do capture the order in which events occur but do not include notions of time; thus, for $\mathsf{T} = \mathsf{acb} \in \mathcal{T}$, we know that a happened before c in T, but we cannot say anything about how much time elapsed between these events.

Definition 2. (Principals) We denote by $\mathcal{P}$ the set of *principals* who participate in the system. Each event has a principal associated with it; intuitively, this is the entity that does the action. (For notational simplicity, we will not explicitly annotate an event with its corresponding principal.) Thus, the principal associated with an event that violates some policy will be the entity that we want to punish; for an event that satisfies the formal definition of a violation below, we will refer to the associated principal as the "violator." (Although our examples below focus on principals as individual humans, our framework does not inherently impose this restriction on violators or principals in general.)

Definition 3. (Outcomes and utilities) Let $\mathcal{T}_{\mathsf{out}}$ be the set containing all traces at which no action is enabled; we call $\mathcal{T}_{\mathsf{out}}$ the set of *outcomes* of the system. For each principal $i \in \mathcal{P}$, there is a *utility function* $\widehat{u}_i : \mathcal{T}_{\mathsf{out}} \longrightarrow \mathbb{R}$.

We will want a notion of an "extended utility" that is defined on all traces and not just on outcomes. Here, we define this as the expected value (w.r.t. some distribution) of a principal's utility function; however, more general approaches (which still capture some sense of depending only on the principal's utility on outcomes) may be of interest.

Definition 4. (Extended utility) We say that a function $f : \mathcal{P} \times \mathcal{T} \times \mathcal{T}_{\mathsf{out}} \longrightarrow \mathbb{R}$ is a *method for extending utilities* if $f(i, \mathsf{T}, \mathsf{T}') \geq 0$ and, for each $i \in \mathcal{P}$ and $\mathsf{T} \in \mathcal{T}$, $\sum_{\mathsf{T}'} f(i, \mathsf{T}, \mathsf{T}') = 1$ (where this sum is take over outcomes T' that extend the trace T). Given such an f and a utility function $\widehat{u}_i$ for principal i, the *extended utility corresponding to* $\widehat{u}_i$ is a function $u_i : \mathcal{T} \longrightarrow \mathbb{R}$ defined by $u_i(\mathsf{T}) = \mathsf{E}_{f(i,\mathsf{T},\mathsf{T}')}[\widehat{u}_i(\mathsf{T}')]$, the expected value of $\widehat{u}_i(\mathsf{T}')$ with respect to the distribution on $\mathcal{T}_{\mathsf{out}}$ defined by $f(i, \mathsf{T}, \cdot)$. We will abuse language and refer to extended utilities simply as "utilities" in contexts where we have established a unique notion of extending each $\widehat{u}_i$ to some u_i.

This view of extended utilities as expected values means that it is natural to allow these functions to be real-valued (instead of, *e.g.*, integer-valued). It thus seems natural to also allow the original utilities to be real-valued as well.

We note that, because outcomes are not extendable, the extended utility of an outcome is the same as the utility of that outcome.

Remark 1. If there is a probability distribution on the outcomes extending each trace, then this naturally gives a method for extending utilities.

Another natural definition for $u_i(\mathsf{T})$ is in terms of the "typical" outcomes that extend T; if there is a "most-typical" outcome that extends a trace, then (in this approach), that would be the only outcome that contributes to the computation of the extended utility of the trace in question. We discuss considerations of typicality below; this approach also influences the notions of causality that we draw upon.

In our definitions of utility, we will generally consider a non-violating trace T_0 that is extended to another trace by a single violation e_v. We will then compare various measures of the violator's utility on different extensions of $\mathsf{T}_0\mathsf{e}_\mathsf{v}$ with measures of his utility on the traces that extend T_0 but do not extend $\mathsf{T}_0\mathsf{e}_\mathsf{v}$. This allows us to formalize, in multiple ways, the intuitive idea that the violator is "worse off" after being punished than if he had not committed the violation.

As considered by Halpern [9] in defining causality, we might have a ranking function ρ on traces whose value indicates how "typical" the trace is; those traces with rank 0 are most typical, those with rank 1 are a bit less typical, *etc.* When using typicality (instead of probabilistic expectation), we will generally be concerned with a principal's utilities in the most typical traces (or in the most typical traces extending a particular trace) or with the utilities as determined by the most typical utility function(s) for that principal. Thus, we will define ranking functions for typicality on both traces and utility functions

Definition 5. (Ranking function for typicality) A *typicality ranking* on a set S is a function $\rho : S \longrightarrow \mathbb{N}$ that assigns to each element of S a non-negative integer. We say that s is more typical than t if $\rho(s) < \rho(t)$. If S is the set $\mathcal{T}$ of traces, we will further assume that for a trace T, each extension T' of T is no more typical than T. It may also be natural to assume (although we do not rely on it here) that T has at least one extension that is no less typical than T.

3.2 Violations

We define violations in terms of a predicate that holds on traces. We assume that once a violation occurs, no further events can undo the fact that the violation occurred; thus, we require that, if the violation predicate holds on a trace, it also holds on all extensions of that trace (so, if the violation predicate holds on ac, then it also holds on acbe). Our focus here is on fundamental definitions for accountability instead of on the particular violations for which principals might be held accountable; in particular, we do not need to formalize time in order to make our definitions, but our framework does not preclude the consideration of violations with temporal aspects (such as violating the requirement "If a transaction involves at least $10,000 in cash, then this must be reported to the IRS within 15 days").

Definition 6. (Violation, violating trace) A *violation predicate* is a predicate that holds on traces such that, if it holds on a trace T, then it also holds on all traces that extend T. If a violation predicate V holds on T, then we say that T is a *violating trace*. If a violation predicate holds on Te but not on T, then we say that the event e is a *violation*. As noted above, the principal associated with a violation may be referred to as the "violator" in the context of the violation.

Remark 2. For policies that require one event to occur within a specified number of events after another event, we think it is natural to assume that these policies refer only to events done by the same principal, *e.g.*, "If principal i does event e_x, then event e_y must be one of the next five events *done by principal* i." This ensures that, for policies such as this, the event defined to be the violation—in this case, the fifth non-e_y event done by i after he did e_x—is done by the same principal who should have done (but failed to do) the second event specified by the policy. In some settings, it may be natural to further restrict the events that are considered in determining policy violations; *e.g.*, the requirement of a prompt response to a message might treat the receipt of a message as an event but then only count non-receipt events in determining promptness.

3.3 Automatic punishment

We start by considering automatic punishment, which is easier to formalize than the more subtle (although perhaps more common) notion of mediated punishment. We define and discuss automatic punishment in two different models. These models differ in how they view expected/typical outcomes. The first model considers the expected utility on outcomes extending a trace (as defined by a distribution on those outcomes); the second model uses the outcomes that are ranked (using a typicality ranking function) as the most typical outcomes extending a trace. In each of these approaches, we define automatic punishment with respect to expected/typical utilities; this is a separate question from distributions on or rankings of traces, and we may again apply these two different approaches to either use a mean utility function or quantify over utility functions that are ranked as "typical" (by a ranking on utilities).

3.3.1 Automatic punishment (probabilistic model)

Working Definition 2. (Automatic punishment (probabilistic model)) Let $\mathsf{T_0 e_v}$ be a violating trace with violation $\mathsf{e_v}$ and associated principal i (*i.e.*, i is the violator). Fix a method f for extending utilities, and let μ be the probability distribution on outcomes obtained by restricting $f(i, \mathsf{T_0}, \cdot)$ (viewed as a unary function of its third argument) to the extensions of $\mathsf{T_0}$ that do not extend $\mathsf{T_0 e_v}$. We say that $\mathsf{e_v}$ is *automatically punished in the probabilistic model* if, for a typical/expected utility $\widehat{u}_i$,

$$\mathsf{E}_{f(i,\mathsf{T_0 e_v},\mathsf{T_{out}})}[\widehat{u}_i(\mathsf{T_{out}})] < \mathsf{E}_{\mu(\mathsf{T'_{out}})}[\widehat{u}_i(\mathsf{T'_{out}})], \qquad (1)$$

where the left-hand side is the expected value of $\widehat{u}_i(\mathsf{T_{out}})$ with respect to the distribution $f(i, \mathsf{T_0 e_v}, \cdot)$ on outcomes $\mathsf{T_{out}}$ extending the trace $\mathsf{T_0 e_v}$, and the right-hand side is the expected value of $\widehat{u}_i(\mathsf{T'_{out}})$ with respect to $\mu(\cdot)$ on outcomes $\mathsf{T'_{out}}$ that extend $\mathsf{T_0}$ but not $\mathsf{T_0 e_v}$.

Remark 3. Working Definition 2 does not impose any restriction on how easy it is to determine whether the Inequality (1) holds. We will not include such restrictions in our fundamental definitions, but they may be a natural part of discussions about, *e.g.*, whether punishment is effective in deterring violations.

This also applies to our other definitions based on inequalities between different utilities.

Example 1. (Second-price auctions) We consider a Vickrey [22] second-price auction with two bidders. Each bidder has a private value for the good being auctioned, and each submits a single bid to an auctioneer, who awards the good to the higher bidder (breaking ties by flipping a coin); the winning bidder is then obligated to pay the bid submitted by the losing bidder. A classic result is that neither bidder can increase her utility (her value of the good obtained—0 for the losing bidder—minus the amount she paid for it) by bidding something other than her true value.

Assume that the policy under consideration is "each bidder should bid her true value for the good" and that each bidder has the standard utility function described above. (This is both the expected utility and the most typical utility.) Assume that each bidder's true values are distributed in some fashion over the discrete range of possible bids, with no value having probability 0. Bidder 1 decides to bid falsely in violation of the policy. Then the expected value (using the probability distribution on outcomes induced by her false bid and the distribution of bidder 2's private value) of her (typical or expected) utility over all the outcomes is strictly less than the expected value of her utility if she had not bid falsely. For both probabilistic and typical notions of utility, Inequality (1) is satisfied and bidder 1 is automatically punished in the probabilistic model.

Remark 4. It is important to note that the automatic punishment in Example 1 does not identify the violator as bidder 1, nor does it even determine that a policy violation took place.

Relationship to incentive compatibility.

Automatic punishment in Example 1 is closely linked to the fact that the Vickrey auction is a truthful mechanism. However, automatic punishment in the probabilistic model does not *require* obedience to be a dominant strategy. Indeed, because we require Inequality (1) to hold only for expected or typical utilities, this is also weaker that asserting that obedience (*i.e.*, not committing a violation) is Bayes–Nash incentive compatible. The following two examples illustrate these distinctions. In each example, the participants have private types—known only to themselves—that determine their utility functions; thus, the utility of principal i on outcome T is given by $\widehat{u}_i(\mathsf{T}, t_i)$, where t_i is i's private type. In the first (Example 2), we use the expected utility (w.r.t. a distribution on the private type of the potential violator). In this case, i has an incentive to commit the violation for one value of his private type but not for the other, even though he is automatically punished for the violation. Although this is weaker than Bayes–Nash incentive compatibility, when expected utility is used, the restriction of Inequality (1) says that, taking the expectation *over the potential violator's private type*, there is no incentive to commit the violation. However, this potential violator will know his own private type before he decides whether to commit the violation; certain principals will have an incentive to commit the violation.[2]

In the second example (Example 3), we use a ranking function to determine which utility functions are the most typical. In this case, i may have an arbitrarily large incentive to violate the policy as long as the utility function that captures this is not ranked as most typical by the ranking function.

Example 2. (Automatic punishment with expected utility) Consider a trace $\mathsf{T_0}$ and a violation $\mathsf{e_v}$ (committed by principal i) that is enabled at $\mathsf{T_0}$. Assume that f is a method for extending utilities, and let μ be the distribution induced by restricting $f(i, \mathsf{T_0}, \cdot)$ to the outcomes that extend $\mathsf{T_0}$ but not $\mathsf{T_0 e_v}$. Assume that there are two possible private types, t_1 and t_2, for i, and that i's utility function depends on his private type (so this will be $\widehat{u}_i(\mathsf{T}, t_j)$ when i has type t_j). Assume that

$$\begin{aligned}
\mathsf{E}_{\mu(\mathsf{T'})}[\widehat{u}_i(\mathsf{T'}, t_1)] &= 1 \\
\mathsf{E}_{\mu(\mathsf{T'})}[\widehat{u}_i(\mathsf{T'}, t_2)] &= 0 \\
\mathsf{E}_{f(i,\mathsf{T_0 e_v},\mathsf{T''})}[\widehat{u}_i(\mathsf{T''}, t_1)] &= 0 \\
\mathsf{E}_{f(i,\mathsf{T_0 e_v},\mathsf{T''})}[\widehat{u}_i(\mathsf{T''}, t_2)] &= \epsilon
\end{aligned}$$

where ϵ is positive, the first two equations involve expected values with respect to $\mu(\cdot)$ as $\mathsf{T'}$ ranges over outcomes extending $\mathsf{T_0}$ but not $\mathsf{T_0 e_v}$, and the second two equations involve expected values with respect to $f(i, \mathsf{T_0 e_v}, \cdot)$ as $\mathsf{T''}$ ranges over outcomes extending $\mathsf{T_0 e_v}$.

Assume that, with probability $1 - q \in (0, 1)$, i's private type is t_1 and, with probability q, i's private type is t_2. If $\epsilon < (1 - q)/q$, then i is automatically punished for committing the violation. However, obedience (not committing the violation) is not Bayes–Nash incentive compatible: Let $s_j(t)$ be a strategy for i, defined in terms of his private type as

[2]While knowledge of the principal's private type might be useful to the designer of an accountability system in order to effectively deter violations, we want the fact that a violator is punished to be independent of his private type.

follows:

$$
\begin{aligned}
s_1(t) &= \begin{cases} \text{Violation if } t = t_1 \\ \text{No violation if } t = t_2 \end{cases} \\
s_2(t) &= \text{Do not commit a violation.} \\
s_3(t) &= \text{Always commit a violation.} \\
s_4(t) &= \begin{cases} \text{No violation if } t = t_1 \\ \text{Violation if } t = t_2 \end{cases}
\end{aligned}
$$

While we would like strategy s_2 to be Bayes–Nash incentive compatible, this is not the case: i's expected utility is higher if he instead chooses s_1.

Example 3. (Automatic punishment with typical utility) Consider again a trace T_0 and a violation e_v (committed by principal i) that is enabled at T_0. Fix a method f for extending utilities, and let μ be the distribution induced by restricting $f(i, \mathsf{T}_0, \cdot)$ to the outcomes that extend T_0 but not $\mathsf{T}_0\mathsf{e}_v$. Assume that i has utility functions $\{\widehat{u^\beta}_i\}_{\beta \in B}$ (the "bad" utilities, writing $\widehat{u^\beta}_i(\mathsf{T})$ for some utility function $\widehat{u}_i(\mathsf{T}, t_{bad})$) and $\{\widehat{u^\gamma}_i\}_{\gamma \in \Gamma}$ (the "good" utilities, using a similar convention); assume that i has an incentive to commit the violation if his utility is "bad" and that he has an incentive not to commit the violation if his utility is "good." Assume also that there is a ranking function that assigns to each of i's possibilities a natural number, with a rank of 0 indicating one of the most typical utilities, *etc.* Regardless of how large his incentive to commit the violation is in the "bad" utilities, as long as the ranking function is such that all of the most typical utilities for i are "good," then Inequality (1) is satisfied.

3.3.2 Automatic punishment (with ranking functions)

For automatic punishment in terms of ranking functions, we assume some ranking on traces that indicates how typical each trace is. We then say that a violation is automatically punished if, for the expected or typical utility of the principal who commits the violation, all of the typical outcomes that extend the violating trace $\mathsf{T}_0\mathsf{e}_v$ yield a lower utility for the violator than all of the typical outcomes that extend T_0 but that do not extend the violating trace $\mathsf{T}_0\mathsf{e}_v$.

It is important to note that, for atypical utilities, the violator could benefit (possibly significantly) from committing the violation e_v.

Working Definition 3. (Automatic punishment (with ranking functions)) Let $\mathsf{T}_0\mathsf{e}_v$ be a violating trace with violation e_v, and let ρ be a typicality ranking on traces. We say that e_v is *automatically punished in terms of the ranking* ρ if

- for the average utility $\widehat{u}_i$ (in the case of average utilities) or for every most-typical utility $\widehat{u}_i$ (in the case of typical utilities),
- for every ρ-most-typical outcome $\mathsf{T}'_{\mathsf{out}}$ extending T_0 but not $\mathsf{T}_0\mathsf{e}_v$,
- and for every ρ-most-typical outcome $\mathsf{T}''_{\mathsf{out}}$ extending $\mathsf{T}_0\mathsf{e}_v$,

we have

$$\widehat{u}_i(\mathsf{T}''_{\mathsf{out}}) < \widehat{u}_i(\mathsf{T}'_{\mathsf{out}}). \tag{2}$$

As with automatic punishment using distributions on outcomes, automatic punishment defined by ranking typical traces is distinct from the incentive compatibility (w.r.t. the utilities $\widehat{u}_i$) of not committing the violation in question.

3.4 Mediated punishment

We say that punishment is *mediated* if it is produced by some event that, in turn, is caused (perhaps through a chain of events) by the fact that e_v is a violation. Unlike automatic punishment, in which the violator's utility is decreased (in an expected or typical sense) immediately upon committing the violation, mediated punishment allows the violator's expected/typical utility to increase sometime after the violation is committed but before the punishment is effected. It is also possible that the violator's expected/typical utility would decrease for unrelated reasons following the violation but before the punishment; we need to make sure that we do not need to completely undo the effects of "good luck" in order to punish a violator, and we do not want "bad luck" to fit the definition of punishment.

In defining mediated punishment, we make use of causality in multiple ways. Halpern [9] has proposed a framework for causality (refining one proposed jointly with Pearl [10]); we do not recapitulate that framework here, but we view that as a natural formal tool to use in conjunction with our definitions.

As discussed below, we may want to have events in our traces that correspond to the fact that a violation has occurred, or to the detection of the violation, that are separate from the violations themselves. (In particular, we will want to treat these as the causes of punishment.) To do this, we may add events of the form $\mathsf{Viol}(\mathsf{e}_v)$; such an event might indicate the fact that e_v is a violation, that this violation is detected, *etc.*

Remark 5. (Events enabled at subtraces) An important aspect of our definitions of mediated punishment is the removal from a trace of the subsequence of events that are causally dependent upon a given event (in particular, a violation). We make the assumption that, if an event e is enabled at a trace and that event is not causally related to an event e_v earlier in the trace, then e would still be enabled if e_v had not happened (and if none of the things causally dependent upon it had not happened). More formally, if T is a trace, e_v is an event in T, and e is an event that is enabled at T and not causally dependent on e_v, then we assume that e is still enabled at the trace T' that is obtained from T by removing e_v and all of the subsequent events in T that are causally dependent upon e_v.

We believe this is a plausible assumption in general; if e does not depend on the removed events, then neither should the fact that it is enabled. When we use this assumption, e_v will be the policy violation and e will be an event that contributes to the good or bad luck of the violator; this assumption allows us to say that the violator could have had the same luck without the violation and then to punish him with respect to the utility of his luck (either good or bad) as it actually played out.

3.4.1 Mediated punishment (probabilistic model)

Working Definition 4. (Mediated punishment (probabilistic model)) Let $\mathsf{T}_0\mathsf{e}_v$ be a violating trace with violation e_v, let the event $\mathsf{Viol}(\mathsf{e}_v)$ capture the fact that e_v is a violation

(committed by principal i), let T extend $\mathsf{T}_0\mathsf{e}_\mathsf{v}$, and let e_p be enabled at T. Let T' be the trace obtained from T by removing e_v and all events that were causally dependent upon it, and let f be a method for extending utilities. We then say that e_v is *punished by the mediating event* e_p *in the probabilistic model* (or that e_p *mediates punishment of* e_v *in the probabilistic model*) if, for a typical/expected utility $\widehat{u}_i$, we have

1. e_p is caused by $\mathsf{Viol}(\mathsf{e}_\mathsf{v})$ (possibly through a causal chain of events)
2. $\mathsf{E}_{f(i,\mathsf{Te}_\mathsf{p},\mathsf{T}''_{\mathsf{out}})}(\widehat{u}_i(\mathsf{T}''_{\mathsf{out}})) < \mathsf{E}_{f(i,\mathsf{T}',\mathsf{T}'_{\mathsf{out}})}(\widehat{u}_i(\mathsf{T}'_{\mathsf{out}}))$

where the left-hand side of Condition 2 is the expected value of $\widehat{u}_i(\mathsf{T}''_{\mathsf{out}})$ with respect to $f(i, \mathsf{Te}_\mathsf{p}, \cdot)$ as $\mathsf{T}''_{\mathsf{out}}$ ranges over outcomes extending Te_p, and the right-hand side is the expected value of $\widehat{u}_i(\mathsf{T}'_{\mathsf{out}})$ with respect to $f(i, \mathsf{T}', \cdot)$ as $\mathsf{T}'_{\mathsf{out}}$ ranges over outcomes extending T'. We then call e_p the *punishing action.*

Condition 1 says that the punishing action depends on the fact that e_v was a policy violation and not just on the fact that the event e_v occurred at some point. We think of causality in the sense of Halpern [9]; to enable the use of that approach, we might (as suggested above) add $\mathsf{Viol}(\mathsf{e}_\mathsf{v})$ to the trace immediately following e_v to treat the fact of the violation (and not just the violating event) as an event in the trace.

Remark 6. In particular contexts, it may be useful to consider punishing actions that are caused not by the fact of the violation but by the determination that the event e_v was a violation (*e.g.*, through the verdict of a jury); this might be another interpretation of $\mathsf{Viol}(\mathsf{e}_\mathsf{v})$ although the event would likely then occur somewhat later in the trace than e_v. The usefulness and implications of such an approach are interesting questions for future work.

Condition 2 prevents "bad luck" from being considered punishment, and it ensures that the violator is punished with respect to whatever "good luck" he actually had instead of with respect to his expected utility just before the violation was committed. Even if the violator's expected utility has decreased due to the events since the violation (*i.e.*, the events that extend $\mathsf{T}_0\mathsf{e}_\mathsf{v}$ to the trace T), his expected utility must be further decreased by the event that counts as the punishing action.

3.4.2 Mediated punishment (with ranking functions)

We now turn to the definition of mediated punishment in terms of ranking functions.

Working Definition 5. (Mediated punishment (with ranking functions)) Let $\mathsf{T}_0\mathsf{e}_\mathsf{v}$ be a violating trace with violation e_v, let the event $\mathsf{Viol}(\mathsf{e}_\mathsf{v})$ capture the fact that e_v is a violation, let T extend $\mathsf{T}_0\mathsf{e}_\mathsf{v}$, and let e_p be enabled at T. Let T' be the trace obtained from T by removing e_v and all events that were causally dependent upon it, and let ρ be a ranking function on the traces of the system. We say that e_v is *punished by the mediating event* e_p *in terms of the ranking* ρ (or that e_p *mediates punishment of* e_v *in terms of the ranking* ρ) if, for a typical/expected utility $\widehat{u}_i$, we have

1. e_p is caused by $\mathsf{Viol}(\mathsf{e}_\mathsf{v})$ (possibly through a causal chain of events)
2. for the average/typical utility $\widehat{u}_i$ of the principal who committed the violation, for every ρ-most-typical outcome $\mathsf{T}''_{\mathsf{out}}$ extending Te_p and every ρ-most-typical outcome $\mathsf{T}'_{\mathsf{out}}$ extending T', we have
$$\widehat{u}_i(\mathsf{T}''_{\mathsf{out}}) < \widehat{u}_i(\mathsf{T}'_{\mathsf{out}}).$$
If we are considering the expected utilities, this must hold for the average utility $\widehat{u}_i$. If we are considering typical utilities, this must hold for every most-typical utility function $\widehat{u}_i$.

4. DISCUSSION AND EXAMPLES

Our working definition of accountability (Working Definition 1) is in terms of punishment, which we have now defined in various models and in both automatic and mediated senses. The question remains of how these definitions can be used in connection with our definition of accountability. One aspect of our working definition involves other agents potentially punishing the violator; this corresponds to mediated punishment. It seems that the most natural way of capturing the idea that other principals could punish the violator is to say that there is a sequence of events that they can carry out (either typically or probabilistically) that produce a punishing trace, regardless of what other events occur in the trace. (That is, events that produce a punishing trace are sequentially enabled at extensions of the violating trace Te_v, even if these are interleaved with unrelated events.)

Our working definition also says that, if i does e_v in violation of a policy, then, with non-zero probability, i could be punished. This assumes a distribution on traces that may not be present; it is thus natural to extend Working Definition 1 to include notions of typical traces (as we did in Working Definitions 3 and 5). Under either approach, it remains open exactly what threshold should be used: should punishment be possible simply with non-zero probability, or do we really want a specific, higher threshold (*e.g.*, a requirement that punishment is possible with probability at least $\frac{1}{2}$)?

Beyond the ways in which our definitions fit together formally, they are also flexible and general enough to apply across a range of situations involving mediated and unmediated punishments and involving identifiability or anonymity of parties involved. We consider some examples to illustrate this.

As shown in Sec. 3.3, our definitions can capture accountability in second-price auctions. In that setting, the punishment is unmediated, and nobody learns the identity of a violator or even whether a violation takes place.

In both the physical world and in computerized systems, accountability is often accomplished using identification. *I.e.*, a system enables the identification of those that have committed violations, and then punishments can be determined appropriately (for example, by a court or other due process). Our definitions can easily be applied to such situations.

A standard example of this in the physical world is given by the judicial system as a method of punishing violations of laws. (We return to this in considering "three strikes" laws below.) In this setting, the judicial system has two major roles: first, determining whether a particular individual has broken a particular law (*i.e.*, committed a violation) and, second, determining the appropriate punishment to set (*e.g.*, jail time and/or fines). Although there are some uncertainties involved to a potential violator, such as whether the

violator will be caught and tried and whether the evidence presented will be conclusive enough to warrant conviction, assuming the system works as expected (*e.g.*, fines and sentencing are typically set high enough to reduce the violator's utility, juries draw the right conclusions, *etc.*), then the punishments are the jail times and/or fines, and accountability is provided by the possibility that these violators will be punished.

Standard examples of computerized systems in which accountability is accomplished using identification include the electronic cash and anonymous-communication protocols [3–6, 21] mentioned in Sec. 1. Of particular interest for our purposes is the DISSENT protocol defined in [6]; the "accountability guarantee" given by DISSENT is that, if one or more participants "disrupt" a protocol execution, then *at least one of the disruptors* will be identified. This guarantee satisfies our requirement for an accountability mechanism, namely that a principal that violates a policy *could be* punished; we do not require that all violators be punished with high probability, and the DISSENT protocol does not, in its current form, satisfy such a requirement. This is a different take on "could be punished" than the one presented in Sec. 2.3, in which a law-enforcement agency could justifiably punish a violator but decides not to, because doing so might compromise an ongoing investigation. In the DISSENT protocol, there is no conscious decision made not to punish a specific violator. Rather, the protocol guarantees that at least one violator will be punished by having his identity exposed, but no one knows *which* subset of the violators will be punished; accountability is achieved, because any principal who chooses to disrupt the protocol knows that he could be among those violators who are punished.

Our definitions can also apply in cases where strong identification is not provided to the participants in a particular transaction, but accountability is achieved based on other existing relationships.

For example, Lampson [14] suggests a method for deterring spam: "reject email unless it is signed by someone you know or comes with 'optional postage' in the form of a link certified by a third party you trust, such as Amazon or the U.S. Postal Service; if you click the link, the sender contributes a dollar to a charity." In this case, the third party may or may not know the identity of the sender, but the recipient need not, and the sender is accountable in any case. In the context of our definitions, we could define the violation to be sending mail to a recipient who considers it spam. The punishment comes from the recipient's clicking on the link costing the sender a dollar, and the accountability comes from the possibility of punishment. Let T be an outcome in which principal i has sent one or more messages that the recipients might have considered spam. If i has sent these messages for a commercial purpose, he will assign a dollar value $w_i(\mathsf{T})$ to each such outcome. If q of the messages in T that i sends are deemed by their recipients to be spam, then $\widehat{u}_i(\mathsf{T}) = w_i(\mathsf{T}) - q$. This expression for i's utility makes clear that a bulk mailer should be willing to pay for quite a bit of "optional postage" if he expects his bulk mailing to result in a very valuable outcome and, conversely, that he could profit from an outcome T with very small $w_i(\mathsf{T})$ if the number of recipients who regard his email as spam could be made correspondingly small.

The usefulness of defining punishment with respect to typical utility functions is perhaps reflected in "three-strikes" laws. In particular, the standard penalties that apply to the first two convictions might be viewed as punishing violators whose utility functions are ranked as most typical (according to a ranking). Those who are not deterred by this—and who continue to commit violations of which they are convicted—might be inferred to have atypical utility functions, and the penalties are recalibrated in an attempt to effectively deter them as well.

For a more precise example, we might consider a population with four different utility functions $\widehat{u^1}$, $\widehat{u^2}$, $\widehat{u^3}$, and $\widehat{u^4}$; these correspond, respectively, to: the majority of the people, who are deterred from committing a violation—*e.g.*, theft—by the prospect of a three-year prison sentence and payment of restitution; a minority of the people who are not deterred from theft by these penalties, but who are deterred by the prospect of a l0-year prison sentence; a much smaller minority of the people who are borderline sociopaths, undeterred by the 10-year prison sentence, but who might be deterred by the prospect of a life sentence; and an extraordinarily small proportion of the population—the sociopaths—who can never be deterred from crime. A person i has the opportunity to commit theft after the events T_0 have occurred; he does so (represented by the event e_v), an investigation ensues, and he is eventually tried, convicted and sentenced to three years in prison and the payment of restitution (with T representing everything—including unrelated events such as i's winning the lottery and being hit by a bus—up to his sentencing, and e_p representing his sentencing). The use of mediated punishment (in the probabilistic model) with typical utilities allows us to capture our intuition that i has been punished, even if he happens to be a sociopath. To see this, we restrict our attention to the case that i's utility is $\widehat{u^1}$; in this case, i expects to be worse off after being sentenced than he would have been had he not committed theft, and if the ensuing investigation, *etc.*, had not taken place (but if he had still won the lottery, been hit by a bus, *etc.*). Formally, we let T' be the trace obtained from Te_p by removing the theft e_v and everything it caused; i's expected utility (assuming his utility function is the most-typical one, namely $\widehat{u^1}$) on the outcomes extending Te_p is less than his expected utility on the outcomes extending T'. This then satisfies our definition of mediated punishment in the probabilistic model when typical utilities are used.[3]

Although the lightest sentence qualifies as punishment, and it deters the majority of the population from committing theft, it does not deter *all* of the population. From the perspective of preventing crime (and not just meeting the definition of punishment), harsher sentences are needed to deter people whose utility function is not $\widehat{u^1}$ from theft; the "third-strike" punishment of life imprisonment deters almost all of the population from theft. This raises the question of what is an *effective* punishment; it suggests that, with respect to typical utilities, one approach would be to view ef-

[3]By contrast, the sociopaths' utility function $\widehat{u^4}$ may be so extreme that the average u—weighted by the corresponding proportion of the population—of $\widehat{u^1}$, $\widehat{u^2}$, $\widehat{u^3}$, and $\widehat{u^4}$ may be such that the expected value of u on the outcomes extending Te_p is greater than its expected value on the outcomes extending T'. This would prevent us from saying—in the expected-utility model—that the sentence of three years in prison plus payment of restitution qualifies as punishment.

fectiveness as being with respect to utilities whose typicality rank is bounded above by some value r.

If, whenever someone commits theft, there is always some non-zero probability that he will be tried, convicted, and sentenced, then—because the penalty in question qualifies as punishment—this satisfies our working definition of accountability. In particular, the thief is *accountable for* his theft (although he might escape punishment in a particular case). As noted below, the use of "accountability" this way may be more general than is desirable. Intuitively, though, this captures the idea that there is some accountability structure in place surrounding this potential crime.

5. CONCLUSIONS AND FUTURE WORK

In our definitions of punishment, we have considered two approaches (expectation and the use of a typicality ranking function) to determine which outcomes and which utility functions to use. As we have argued above, in at least some settings our intuition for punishment is captured by considering the utility functions that are ranked most typical. As noted by Halpern [9], there are other approaches to typicality in the literature; it would be interesting to explore definitions of punishment (and other notions related to accountability) in terms of those approaches.

In modeling "accountability," we have striven for generality. Our resulting framework encompasses everything that may be more effectively handled with "deterrence," as Lampson [14] uses the term, than with the security-research community's preferred approach, namely prevention. This very general use of the word "accountability" may create barriers to adoption, because the word connotes "answerability" and "standing up to be counted" in ways that suggest formal adjudication and the inability to act anonymously while remaining accountable. We explore these terminological issues more fully in a recent paper with Hendler and Weitzner [7]. Here, we stress that the value of our formal framework does not depend on the use of the word "accountability." If it were considered a framework for "deterrability," and the word "accountability" were reserved for scenarios involving mediated punishment in which all actors are identifiable, all of our definitions would remain meaningful and applicable.

There are numerous other notions related to accountability. These include:

Compensation This complements punishment. If a principal is the victim of a policy violation (such as a theft), he might be compensated. Note that while "victimless" crimes still naturally lead to punishment and can easily be captured in our system, the corresponding notion ("perpetratorless crime") does not seem to fit as naturally into intuitive understandings of accountability.

Detection/Diagnostics If (mediated) punishment is to be carried out in response to a violation, there are numerous questions surrounding detection and system diagnostics that are important to answer.

Authorization There are many different (often domain-specific) notions of authorization, and these often play an important role in defining policy violations.

Legal theories of liability might also inform the further development of accountability.

From a technical perspective, there are parallels between our various different definitions of punishment that suggest there may be a common generalization. Such an approach may be both technically useful and helpful in better understanding accountability and related concepts.

6. ACKNOWLEDGMENTS

We appreciate helpful feedback from Steve Greenwald, Joe Wegehaupt, Hongda Xiao, and the NSPW 2011 participants.

7. REFERENCES

[1] D. Andersen *et al.*, "Accountable Internet Protocol (AIP)," in *Proceedings of the 32nd ACM SIGCOMM Conference*, 2008, pp. 339–350.

[2] G. Bella and L. Paulson, "Accountability protocols: formalized and verified," *ACM Transactions on Information and System Security*, vol. 9, no. 2, 2006, pp. 138–161.

[3] J. Camenisch and A. Lysyanskaya, "An efficient system for non-transferable anonymous credentials with optional anonymity revocation," in *Proceedings of EUROCRYPT '01*, Lecture Notes in Computer Science, vol. 2045, Springer, Berlin, 2001, pp. 93–118.

[4] J. Camenisch, A. Lysyanskaya, and M. Meyerovich, "Endorsed e-cash," in *Proceedings of the 28th IEEE Symposium on Security and Privacy*, 2007, pp. 101–115.

[5] D. Chaum, "Blind signatures for untraceable payments," in **CRYPTO '82**, Plenum Press, New York, 1982, pp. 199–203.

[6] H. Corrigan-Gibbs and B. Ford, "Dissent: accountable anonymous group messaging," in *Proceedings of the 17th ACM Conference on Computer and Communication Security*, 2010, pp. 340–350.

[7] J. Feigenbaum *et al.*, "Accountability and Deterrence in Online Life (Extended Abstract)," in *Proceedings of the 3rd International Conference on Web Science*, ACM, 2011.

[8] R. Grant and R. Keohane, "Accountability and Abuses of Power in World Politics," *American Political Science Review*, vol. 99, no. 1, 2005, pp. 29–43.

[9] J. Halpern, "Defaults and normality in causal structures," in *Proceedings of the 11th Conference on Principles of Knowledge Representation and Reasoning*, 2008, pp. 198–208.

[10] J. Halpern and J. Pearl, "Causes and explanations: A structural-model approach–part I: Causes," *British Journal for the Philosophy of Science*, vol. 56, no. 4, 2005, pp. 843–887.

[11] R. Jagadeesan *et al.*, "Towards a theory of accountability and audit," in *Proceedings of the 14th European Symposium on Research in Computer Security*, Lecture Notes in Computer Science, vol. 5789, Springer, Berlin, 2009, pp. 152–167.

[12] R. Küsters, T. Truderung, and A. Vogt, "Accountability: definition and relationship to verifiability," in *Proceedings of the 17th ACM Conference on Computer and Communications Security*, 2010, pp. 526–535.

[13] B. Lampson, Notes for presentation entitled "Accountability and Freedom," 2005.

http://research.microsoft.com/en-us/um/people/blampson/slides/AccountabilityAndFreedom.ppt.

[14] B. Lampson, "Usable Security: How to Get it," *Communications of the ACM*, vol. 52, no. 11, 2009, pp. 25–27.

[15] J. Mashaw, "Structuring a 'Dense Complexity': Accountability and the Project of Administrative Law," Article 4 in *Issues in Legal Scholarship: The Reformation of American Administrative Law*, 2005, http://www.bepress.com/ils/iss6/art4.

[16] MIT Decentralized Information Group, webpage on Social Web Privacy, accessed July 29, 2011. http://dig.csail.mit.edu/2009/SocialWebPrivacy/

[17] MIT Decentralized Information Group, Theory and Practice of Accountable Systems project homepgae, accessed July 29, 2011. http://dig.csail.mit.edu/2009/NSF-TPAS/index.html

[18] MIT Decentralized Information Group, Transparent Accountable Datamining Initiative project homepage, accessed July 29, 2011. http://dig.csail.mit.edu/TAMI/

[19] R. Mulgan, **Holding Power to Account: Accountability in Modern Democracies**, Palgrave MacMillan, Basingstoke, 2003.

[20] H. Nissenbaum, "Accountability in a Computerized Society," *Science and Engineering Ethics*, vol. 2, no. 1, 1996, pp. 25–42.

[21] P. Tsang *et al.*, "Blacklistable anonymous credentials: blocking misbehaving users without TTPs," in *Proceedings of the 14th ACM Conference on Computer and Communication Security*, 2007, pp. 72–81.

[22] W. Vickrey, "Counterspeculation, auctions, and competitive sealed tenders," *Journal of Finance*, vol. 16, no. 1, 1961, pp. 8–37.

[23] D. Weitzner *et al.*, "Information Accountability," *Communications of the ACM*, vol. 51, no. 6, 2008, pp. 82–88.