Accountability as an Interface between Cybersecurity and Social Science*

Joan Feigenbaum† Aaron D. Jaggard‡ Rebecca N. Wright§

Overview "Accountability" is generally agreed to be important, although this term is used to mean many different things. Across these various uses, it is related to the idea of deterrence instead of prevention. Computer-science-based work on accountability is providing models to formalize accountability and disambiguate it from related notions. This work also raises many questions that need to be informed by social science, suggesting the potential for fruitful interaction at this cross-disciplinary interface.

The benefit of computer-science perspectives We have proposed a formal view of "accountability" in trace-based and game-theoretic terms. This framework helps to make precise different notions related to "accountability" and to distinguish between them. This and other frameworks also open up the possibility of proving formal relationships (*e.g.*, implications and tradeoffs) involving different accountability-related notions. Interaction with the social sciences will enrich formal models of accountability and make them more realistic. In turn, these enhanced models will help answer real-world social-science questions related to accountability.

The need for social-science perspectives Considering formal models of accountability in computer science, we may identify issues for which we expect social-science perspectives to make significant contributions. Some of these are below. However, increased interaction between computer and social scientists may change the framing of these questions. More fundamentally, interactions between computer and social scientists would better identify the *types* of questions that should be studied at this cross-disciplinary interface.

In studying deterrence, the question of what constitutes "effective deterrence" is an important one: bad behavior should *actually be deterred*, which may be somewhat independent of whether any particular technical definition is satisfied. An understanding of human responses to a range of incentives, both positive and negative (such as payments, incarceration, shame, and praise) would inform both more complete models for further study as well as more effective systems for real-world use.

Relatedly, the utility functions studied in connection with accountability may differ dramatically between people. One utility function may be most typical in a population, another one slightly less typical, and so on. Effectively deterring undesired behavior by a "typical" individual may be relatively easy using socially acceptable means. By contrast, deterring even the sociopaths, without knowing in advance who they are, might require measures that would be draconian if applied uniformly. This leads naturally to the example of "three strikes" laws, which suggests that it would be beneficial to further explore formal frameworks for accountability in connection with a broader and deeper knowledge of criminology, sociology, law, and other disciplines.

Accountability often makes use of causality. Our model uses causality to connect punishments to violations; as noted by various people, causality also arises in treating "blame" as something other than a black box. Additionally, identifying violations and assigning blame often make use of some sort of evidence. Causality and evidence have been studied, *e.g.*, at the boundary of computer science and philosophy. Further work on these concepts in the context of accountability would help ensure that accountability systems are viewed as legitimate (for example, that punishment is meted out only when it is sufficiently justified).

"Accountability," "deterrence," and related terms have various connotations in colloquial usage. For example, our candidate definition of accountability is in terms of punishment and not "calling to account," *etc.* Should we instead use a term like "deterrence" for what is captured by our candidate definition of "accountability?" Our formal framework also intentionally allows for a violator to be automatically punished, without necessarily identifying the violator or even revealing that a violation occurred. As Weitzner has asked, is it better to reserve the use of "accountability" for cases in which someone *knows* that a violation occurred? Even once accountability-related terms are sufficiently disambiguated for interdisciplinary communication, properly framing them will be important for communicating to broader audiences (such as end users or consumers) what accountability systems are intended to do as well as what they do not do.

*Partially supported by NSF grants CNS-1016875 and CNS-1018557.

†Department of Computer Science, Yale University. joan.feigenbaum@yale.edu

‡Formal Methods Section (Code 5543), U.S. Naval Research Laboratory. aaron.jaggard@nrl.navy.mil

§DIMACS and Department of Computer Science, Rutgers University. rebecca.wright@rutgers.edu