Microsoft

Establishing End to End Trust

By Scott Charney*
Corporate Vice President
Trustworthy Computing
Microsoft Corp.

* This paper benefited from the many reviewers who provided substantive comments and helped to shape this paper. Please see Appendix A for an admittedly incomplete list of contributors.

Contents

I. Introduction

It is not an overstatement to say that the Internet has transformed the way we live. Social networking represents the new town square; blogging has turned citizens into journalists, and e-commerce sites have spurred global competition in the marketplace. But along with the Internet's phenomenal growth has been a growth in computer-related crimes. The range of criminal activity that the Internet supports is vast—from consumer threats (e.g., becoming a "bot," ID theft, and child endangerment), to enterprise threats (e.g., the theft of stored personally identifiable information, economic espionage, and extortion via threats of denial of service attacks), to government threats (e.g., information warfare), there is little doubt that creative, adaptive, and sophisticated adversaries are misusing the Internet to bad effect. According to Gartner, "Phishing attacks in the United States soared in 2007 as $3.2 billion was lost to these attacks. The survey found that 3.6 million adults lost money in phishing attacks in the 12 months ending in August 2007, as compared with the 2.3 million who did so the year before."[1] Governments have expressed increasing concern about public safety and national security, including information warfare. Indeed, if we want online activity to provide all its potential benefits, security on the Internet cannot remain at current levels.

At the same time, there is increasing concern about privacy in the digital age. Some of this concern is a function of the criminal landscape since some crimes target personally identifiable information for theft and misuse. But privacy concerns are far broader than just public safety concerns; there are additional non-crime-related concerns about the amount of data we generate in our electronic lives and how that data can be collected, aggregated, analyzed, disseminated, and used. Although this is a very important issue, the focus of this paper is security and the privacy benefits that accrue from efforts to combat cybercrime by, in part, giving users better control over their digital personas. A fuller discussion of broader privacy issues must be left for another day.

Although security, and not privacy, is the focus of this paper, it is important to understand the interplay between the two realms. The primary goal of security is to protect the confidentiality, integrity, and availability of data and systems—the attributes that criminals attack. To the extent security protects the confidentiality of data, it serves to protect privacy. But security often involves collecting evidence of a person's activities (both evidence of past activity, such as audit logs; and ongoing activities, such as keystroke monitors). In that context, security may involve surveillance and raises serious privacy concerns, a point that must be kept in mind when one addresses the growing cybercrime problem.

Our initial focus on security is appropriate because the Internet has proven to be a great medium for committing crime. Not only was it designed without concern for security (the initial "users group" were trusted people working for, or funded by, the United States government), but it has four key attributes

[1] Gartner, Inc. "Gartner Survey Shows Phishing Attacks Escalated in 2007; More than $3 Billion Lost to These Attacks," Dec. 17, 2007.

The U.S. Department of Justice Bureau of Justice Statistics and the U.S. Department of Homeland Security National Cyber Security Division conducted a National Computer Security Survey to provide official national statistics on computer security incidents across industry, but the results have not yet been released. See http://www.ojp.usdoj.gov/bjs/survey/ncss/ncss.htm.

The effect of these attacks may be a diminished use of information technology. For example, although online banking continues to grow, some believe that fraud worries will cut 1 to 2 percent of online banking growth. Additionally, research shows that consumers who bank online engage in fewer online activities than they would if they felt safer. See "Click! Online banking usage soars," at http://www.msnbc.msn.com/id/6936297/.

that attackers love: (1) global connectivity; (2) anonymity[2]; (3) lack of traceability; and (4) valuable targets. In addition, it is difficult for computer users to know, or find out, what programs are running on their machines, what machines they are connecting to, and with whom they are dealing. As a result, those prone to prey electronically on others have considerable opportunity for success, with little risk of being identified and being held accountable for their actions.

It is correct to both assume and hope that the use of the Internet will continue to grow, expanding its reach and resulting in even more online activity. Indeed, new connection models, such as "Anywhere Access" (where peer-to-peer connections enable new business models and allow people to access their data from anywhere on any device), mean that global connectivity and the number of valuable targets will increase, thus attracting even more criminal activity. It is therefore critically important that we find a way to both improve the security of computer networks and put people back in control of their computer environment. Although Microsoft Corp. and many other organizations have taken significant steps to improve the security and privacy of their products and services in this increasingly connected world (the Microsoft path is described in more detail below), these activities alone will not make the Internet secure enough and privacy-enhanced enough for many of its potential uses. Thus, this paper is an invitation to discuss how one might fundamentally "change the game," and provides a framework for discussing the myriad of social, political, economic, and technological issues that must be addressed if we want to create a meaningfully more secure and privacy-enhanced Internet.

In short, in our view changing the game requires two things: (1) building a "trusted stack," with suitably strong authentication of hardware, software, people, and data; and (2) improving the ability to audit events to provide accountability. We must also grant people better control over their digital personas to enhance privacy. This trusted stack, combined with better mechanisms to protect privacy, will enable End to End (E2E) Trust—giving people, devices, and software the ability to make and implement good decisions about who and what to trust throughout the ecosystem. This will help protect security and privacy as well as help bring criminals to justice when electronic malfeasance occurs. In sum, the opportunity exists to create a trusted, privacy-enhanced Internet.

II. Microsoft Security Strategy

In January 2002, Bill Gates announced the Trustworthy Computing initiative at Microsoft. Developed in the summer of 2001, the initiative focused on building trust in the IT ecosystem. Based on the success of the telephone, it was clear that people needed technology to have certain core attributes if it was to be embedded fully in their lives, including reliability, security, and protections for privacy. Although TwC, as it became known, was designed to focus on all these attributes, most people initially equated TwC with security. This occurred because the announcement of TwC post-dated the

[2] It is difficult to speak of "anonymity" and "identity" in absolute terms; indeed, the "anonymous" and "identified" modes commonly experienced online today are functional but not absolute.

A truly anonymous state is difficult to achieve because it is theoretically possible to link actions back to specific computers if enough parties collude. Similarly, a truly identified state is also difficult to achieve because credentials can be shared, stolen, or otherwise compromised, making it difficult to know who is actually taking actions online.

Between these two absolute poles lies the broad range of states including "functional anonymity," "functional identity," and the range of states in between. Moreover, as I discuss in this paper, anonymity is an important social value, and should be preserved and enhanced through both technology and social policy, so that, in the right situations, people are able to choose whether they want to be anonymous or identified.

events of 9/11, an event that reshaped computer security by raising new concerns about critical infrastructure protection.

The terrorist attacks of Sept. 11, 2001, which shut down the U.S. stock market for five days, refocused government and industry attention on the security of the information infrastructure. Shortly thereafter, the Nimda and Code Red worms broadly affected Microsoft products and customers, and Microsoft redirected considerable energy and resources to improving the security of its products. In a company where mandatory rules are frowned upon because of their adverse impact on innovation and the "tax" they impose on staff, the Security Development Lifecycle (SDL) was declared mandatory,[3] and products that failed to pass a Final Security Review were subject to "stop ship" orders until issues were escalated and resolved.[4] As a result of these efforts, the security of Microsoft products improved (vulnerability counts for products have dropped continuously since the SDL's implementation), and although security remains a major issue for customers, Microsoft reputation for security has improved.

Despite these advancements, critics complained that Microsoft lacked a comprehensive, holistic "security strategy."[5] This is not entirely fair. In fact, a strategy was created and implemented, and it has evolved over time. It started with SD3, Microsoft shorthand for "Secure by Design, Secure by Default, and Secure in Deployment." Simply put, if software were more securely designed (which, in this context, includes both design and development), placed in the market in a more secure state (by, for example, reducing attack surface by turning features off by default and having users run as "standard users" as opposed to "system administrators"), and maintained securely once in the market (through, for example, a robust response process that included both timely upgrades and the tools to deploy them quickly), the IT environment would be safer. There was, in fact, nothing wrong with this strategy as a foundation, and SD3 remains important today.

The problem with SD3 lies in its inherent limitations. Even if products are engineered to be "Secure by Design" and vulnerability counts continue to drop, it is indisputable that the number of vulnerabilities in large and complex products (several of which are likely to be installed on a single system) cannot be reduced to zero in the foreseeable future. "Secure by Default" is inherently limited because the attack surface can only be reduced, not eliminated, and features are created precisely because a broad set of users need the feature activated. Similarly, many legacy software applications require the user to run as "admin," thus undermining some of the intended security benefits of running as a standard user. And although "Secure in Deployment" is important, patches are reverse engineered, and exploits launched, faster than many users can test and deploy patches. In addition, there is an increasing number of zero-day exploits that offer no opportunity for the deployment of a patch in the first instance. Finally, SD3 is focused primarily on product security, and while some of its elements,

[3] The Security Development Lifecycle is a process for developing software that includes security milestones (such as the development of threat models and the use of code checking tools) throughout the product life cycle. The SDL is mandatory for software that connects to the Internet, is used in enterprises, or can be used to store or transmit personal information. For more details, see Howard and Lipner, "The Security Development Lifecycle," Microsoft Press®, 2006.

[4] The Final Security Review involves an independent (non-product team) group of security experts evaluating adherence to the SDL. The question they are tasked to answer is, "From a security perspective, is the product ready to ship?" In practical terms, the goal is to ship products with no known critical or important vulnerabilities.

[5] Part of the problem is that Microsoft security strategy is often discussed in relation to products recently released or soon coming to the market. Although products may implement elements of a security strategy, they cannot be the sine qua non for it, especially since products are built to respond to the needs of the market and the market has often favored functionality over security.

particularly Secure in Deployment, include security manageability, SD3 was not intended to, and cannot, fully address the vulnerabilities and exposures that come from managing multiple complex software products in heterogeneous environments, especially when considering the full life cycle of deployment, operations, maintenance, upgrade, and destruction.

Recognizing some of these limitations, SD3 was supplemented by Defense-in-Depth. Defense-in-Depth recognizes that some security efforts will fail and others might provide a separate layer of protection. By way of example, notwithstanding the fact that Microsoft has (1) reduced vulnerabilities in code; (2) turned features off by default; (3) turned on a firewall by default in Service Pack 2 for Windows® XP and then in Windows Vista®; (4) instructed users to run (and later provided) anti-virus products; and (5) educated users on the risks of running code of unknown provenance, it remains true that users will click on malicious attachments sent to them from unknown sources. Thus, Microsoft also provides the Malicious Software Removal Tool (MSRT) to clean machines that have been infected.

There remained, however, other more specific threats not well addressed by SD3 or Defense-in-Depth. For example, spam does not normally exploit vulnerabilities, nor would one turn off mail by default. There is also very little a specific user or enterprise can do to prevent a distributed denial-of-service attack from a botnet. As a result, Microsoft started working on threat mitigations for specific issues. With regard to phishing and spam, for example, it engaged in broad consumer education campaigns and worked on developing technological solutions such as phishing filters and SenderID. For both phishing and botnets, Microsoft began working more extensively with law enforcement to identify phishers and botnet herders in an attempt to create deterrent to such activity, even though the deterrent effect is limited by the current environment because it is hard to find offenders, and criminal penalties may be applied without sufficient force.

III. Evolving the Security Strategy: Creation of the Trusted Stack and Enabling End to End Trust

Despite SD3, Defense-in-Depth, and specific threat mitigation, security and privacy remain major concerns of computer users. As people look to engage in an increasing number of personal and commercial activities online, it becomes important to address their growing demands for both security and privacy. Part of the problem is that the security solutions employed to date are primarily defensive technical measures that, while effective in mitigating particular avenues of attack, do not address an adversary who is adaptive and creative and will rapidly shift tactics. Thus, for example, hardening of the operating system caused attackers to move “up the stack” and attack applications, as well as refine social engineering techniques that technology today is ill-equipped to help prevent.

This is not to say that these initial security strategies did not and do not have their place; to the contrary, they were foundational, effective in mitigating some serious threats and must continue to be pursued with vigor. Whether it involves creating better static code analysis tools, or more fundamental changes such as using virtualization to sandbox certain activities such as browsing Web sites or downloading code of unknown provenance, the IT industry must continue to do the fundamentals well. But staying the current course will not be sufficient; the real issue is that the current strategy does not address effectively the most important issue: a globally connected, anonymous, untraceable

Internet with rich targets is a magnet for criminal activity—criminal activity that is undeterred due to a lack of accountability. Moreover, the Internet also fails to provide the information necessary to permit lawful computer users to know whether the people they are dealing with, the programs they are running, the devices they are connecting to, or the packets they are accepting, are to be trusted. Thus, one of the most sacred pieces of advice in the physical world—caveat emptor ("let the buyer beware")—is unhelpful.

Determining trust on the Internet is, of course, a very complicated matter. A single trust decision may require a user to consider, simultaneously, whether to trust the device, person, software, and data[6] involved in the transaction. To the extent a user bases his or her decision on an attribute of a component (e.g., the identity of a person or device), there remains a large degree of risk because the Internet makes it easy to provide false identity.

Even assuming identity or an identity claim is truly represented, making accurate trust decisions regarding a person or thing may still be difficult. First, trust is not binary; some things may be trusted completely, trusted not at all, or trusted only for limited purpose.

Second, trust decisions may be static or dynamic. For example, in the physical world, once trust is established, we often maintain that trust for long periods of time until some event shakes our faith. This is often true regarding our trust in people or our trust in the reliability of our mechanical devices (e.g., cars, coffee makers). By contrast, on the Internet, trust may be incredibly dynamic: my computer may be fully patched one minute, missing an update the next, then re-updated quickly and yet still vulnerable to a zero-day exploit. In such an environment, it may be difficult to decide whether a machine should be "trusted" and be allowed to access other network resources.

Third, trust decisions may be influenced by actual or perceived concerns about the impact of an erroneous decision, such as whether a program might be sandboxed or whether a transaction can be rolled back if erroneous.

Fourth, it may be hard to make reliable trust decisions because of the opaqueness of the Internet and its components. In the physical world, people get visual clues from others that we instinctually recognize and assess correctly (a warm smile). Applying these physical world concepts to electronic transactions is difficult. Some of those physical clues are simply unavailable, and electronic equivalents for such clues (authentication and signatures) have proven to be incomplete, difficult to manage, and hard for users to comprehend and act upon. In the absence of usable information, users sometimes consider inappropriate factors. For example, a computer user may take comfort from the fact that a piece of software has been downloaded by a large number of other users, but this may be a statement about popularity as opposed to security. If it turns out that software was not to be trusted and consisted of a keystroke logger that steals personal information, there may be nothing that can be done to rectify the situation short of removing the offending software after the damage has already been done. All of this suggests that creating a more trustworthy Internet depends critically on the realization of vastly improved "trust user experiences" that will communicate in understandable form the significance of certificates, signatures, identities, and access decisions.

[6] A wide range of zero-day attacks highlights the importance of trustworthy data attachments. But it is important to note that the reference here is to source and integrity (that is, who signed it and whether it was altered since being signed). In the security context, "trust" is not meant to include data quality (e.g., whether a particular number in a spreadsheet is accurate).

Although trust may be a complex issue, this does not alter the fact that certain foundational elements must be in place to create a more trustworthy environment. The most important element is an authenticated identity claim (e.g., name, age, or citizenship); absent the ability to authenticate a person (or a personal attribute), machine, software, and/or data—and absent the ability to combine that authenticated data with other trust information (e.g., prior experience, reputation), effective trust decisions cannot be made. Second, absent the ability to identify and prove the source of misconduct, there can be no effective deterrent—no effective law enforcement response to cybercrime and no meaningful political response to address international issues relating to cyberabuse. To date, the "response" to computer abuse of all types has been to increase defenses, but the history of computer security shows that offense will beat defense in cyberspace because attackers have an abundance of time and resources, and may only need to find one weakness, whereas a defender must cover all avenues of attack.[7] Experience shows that most cybercriminal schemes are successful because people, machines, software, and data are not well authenticated and this fact, combined with the lack of auditing and traceability, means that criminals will neither be deterred at the outset nor held accountable after the fact. Thus the answer must lie in better authentication that allows a fundamentally more trustworthy Internet and audit that introduces real accountability.

We must create an environment where reasonable and effective trust decisions can be made. We must also create an environment where accountability—and therefore deterrence—can be achieved. To do this, one must have access to a trusted stack: (1) security rooted in the hardware; (2) a trusted operating system; (3) trusted applications; (4) trusted people; and (5) trusted data. The entire stack must be trustworthy because these layers can be interdependent, and a failure in any can undermine the security provided by the other layers; for example, a document may be created by an identified individual, using secure hardware and a secure operating system, and sent to another as a signed attachment with integrity, but if it was created with an insecure application, it may not be trustworthy. And when trust is misplaced, it must be possible to identify the improvidently relied-upon party and have the right social and political mechanisms in place so that proactive and reactive steps can be taken. An appropriate audit capability can provide the evidence needed to inform response and drive an accountability framework.

IV. Some Initial Thoughts on Identity

Although the issue of identifiers (or identity claims) will be discussed in more detail later, experience has shown that the mere mention of "identity" can result in deep misunderstandings. Thus, before discussing the advantages of a better authenticated and audited infrastructure, some initial comments regarding identity are in order.

[7] This is not a call to develop an offensive cyber-capability, for the effectiveness of retaliation is debatable. One issue relates to targeting; for example, a victim may disable an attacking botnet but the machines "disabled" may actually be tainted consumer machines and the botnet can be reconstituted quickly by the adversary. Second, a "mutual destruction" philosophy assumes equal dependence on technology. Put another way, if an IT-dependent society "shoots" at a less IT-dependent society and both are disabled, it is not clear that the impact is equal. Finally, it is not even clear whether offensive cyber activities are subject to physical world rules. In the physical world, "theatre-level rules of engagement, collateral damage estimation, and positive identification all must be observed before any strike takes place. Rules such as these keep responses proportionate to the political-military goals of an operation…. With cyberspace operation, that framework is not so prominent." "The Dogs of Web War," Air Force Magazine, January 2008.

First, nothing in this paper is meant to suggest that anonymity on the Internet be abolished. To the contrary, anonymity should be preserved and enhanced through both technology and social policy. More important, in the right situations, people should be able to choose whether they want to be anonymous or identified (in whole or in part), and for what purpose. User choice is important.

Second, nothing in this paper is meant to create unique, national identifiers, even if some countries are creating identity systems that do so. Indeed, people often have many identities (e.g., work identities, personal identities, pseudonyms, and temporary or anonymous "identities") and should be able to choose what identity to use in a given situation. More important, people should have the ability to pass identity claims. For example, if one wants to visit sites with content not appropriate for children, individuals should be able to prove age *without* necessarily providing other information about identity. Similarly, if one wants to purchase goods online, it should be possible to pay for that transaction through services that do not require the disclosure of a credit card number. Indeed, focusing on privacy may actually enable new, privacy-centric business models; for example, it may be possible to engage in targeted anonymous advertising since it may be possible to "know" something about someone without knowing who they are. In sum, we should be able to enable more and safer social and commercial opportunities without diminishing privacy by having numerous identities and limiting the personal data we share, thus making data aggregation and analysis more difficult, by design.

Third, nothing in this paper supports the creation of mega-databases that collect personal information. In addition to the fact that people should be able to choose what identity or identity claim to use when, auditing should remain, as it is today, distributed. Further, audit information should be better protected than is usually the case today.

Fourth, there is no claim that creating an authenticated, audited environment has *no* impact on privacy. Privacy is not a state, but a continuum: on one end, there may be anonymous people and no tracking; at the other end, there may be situations where authenticated people act in a highly monitored environment (e.g., airport travelers). People already engage in many non-anonymous activities, such as shopping with a credit card, conducting online banking, and even putting truthful profiles on social networking sites. That said, in the absence of the proper controls, "adding" robust identity in any circumstance where it does not now exist moves us along the continuum. But in a world with new and significant security and privacy risks, we need to provide people and governments with more choices to better manage and address those risks.

Fifth, any system can be abused and, if the risk of serious abuse is significant enough, we might eschew the approach. But the argument that an authenticated ecosystem will inevitably be subject to abuse of such magnitude oversells itself. As noted below, there are historic, economic, social, and political forces that suggest a well-constructed regime is better than none at all, especially in light of the challenges we face on the Internet and the desire of people to be more secure in their daily lives.

Finally, universal buy-in and implementation is not necessary to achieve a modicum of success. The problem today is that those who want greater safety have few effective ways to achieve it. The goal is to provide enough meaningful data to empower people to make trust decisions, even if that data does not exist in all circumstances. In those circumstances where such data does not exist, people should be empowered to disengage or knowingly accept the increased risk.

V. The Benefits of a Trusted Stack

What benefits arise from the fact that people, devices, software, and data are more robustly authenticated and their activities audited? In a general sense, the most obvious benefit of authentication is that it empowers better trust decisions. Auditing creates a better ability to hold people accountable for misconduct, and thereby deter such conduct, assuming that domestic cybercrime laws and international cooperation mechanisms are sufficient. Enabling better trust decisions and accountability will solve specific real-world problems. For example, a well-audited transaction between two authenticated parties serves to protect both sides of the transaction. A bank could more easily authenticate a customer's identity, a customer would have greater assurance that the Web site that he or she was visiting was that of the bank, and both parties could determine what truly happened if any issue arose. By conducting device-to-device authentication, organizations could reduce the number of external hackers with access to their systems, in large part because a hacker would need access to an "authorized" machine to connect to the victim's network. In addition, if an unauthorized access were to occur and better auditing records proved what happened, it would become much easier to apply physical world mechanisms (e.g., law enforcement, political forces) to address cybercrime, economic espionage, and information warfare. Because these mechanisms enable more effective trust decisions to be made throughout the ecosystem—by and about people, devices, software, and data—we call this End to End Trust.

Improved authentication and audit capabilities would generate a host of other opportunities, especially if robust management tools permitted system administrators to increase the amount (or change the type) of audit data collected, depending on the threat level. This helps to balance the need for evidence with the cost of collecting and storing data. The ability to reliably detect and attribute flooding and probing attacks would be increased. Autonomous defense would be possible if, for example, packets likely to be malicious (because they are reliably identified as coming from a dangerous source) could be dropped shortly after entering the network or at a computer's interface to the network. Even the intractable insider threat could be more successfully addressed because better audit tools would make it easier to identify suspicious access patterns for employees in a timely manner.

The authentication of identity, device (and its state), software, and data could be used to generate trust measurements that could also be used to reduce risk to the ecosystem. For example, one of the reasons that large enterprises manage risk relatively well is that they have dedicated IT staff implementing risk management programs. Yet there is no chief information officer for the public, and no mechanism for protecting the broader Internet by taking best practices from enterprises, such as Network Access Protection, and applying those practices to the public. With better authentication and audit, dynamic trust decisions could be made (based upon, for example, the state of a machine) and Internet service providers could use network access controls to limit the activities of "untrustworthy" machines until they were updated.

VI. Setting Reasonable Security Goals

All security strategies, whether designed to ensure physical security or information security, must be based on sound risk management principles. Put more bluntly, it is about risk management, not risk elimination. Both home security and car security (which often includes simply locking windows and doors and, less frequently, installing alarm systems) can be defeated easily; the real question is whether the level of safety offered is reasonable under the circumstances presented.[8] In addition, any security strategy must include an ecosystem strategy and product, and/or service strategy that maps to it; home and car alarms are not valuable without neighbors and/or police who can and will respond.

As the reader looks at some of the ideas below, he or she may quickly identify how certain elements of the security strategy can be defeated. By way of example, a key part of the strategy involves in-person proofing as a condition precedent to creating certain digital identities. One example of this strategy in action might be schools identifying their students and issuing them digital credentials so that students can visit hosted electronic playgrounds where all the other members are children of similar age. Certainly it is true that a worker at the school might be able to get a certificate and lurk on the site. But that worker can also lurk in the schoolyard and potentially cause injury to a child. We cannot completely solve safety problems in the physical world (although we can mitigate the risk through background checks, ensuring multiple adults act as supervisors in the playground, etc.) and we cannot completely solve the problem in the cyber world. But the more important point is that the electronic playground is no longer open to the entire planet and anyone who chooses to self-certify as a child. Equally important, to the extent risk of apprehension creates a real deterrent, this regime provides better opportunities for law enforcement. If an adult does get hold of a child certificate that was properly protected (not open to the world), then the range of suspects is reduced from “anyone in the world” to “anyone who could have corrupted the proofing process or accessed the identity of one of the students.” Although that may still be a non-trivial list (it will include anyone with access to the improperly used certificate), it is still a much more workable case from a law enforcement perspective, especially if future communications from a suspect or to a victim are traceable, or if law enforcement works promptly to secure data under existing authorities (e.g., quickly freezes the records related to the IP address from which the adult connected to the electronic playground).

In short, it is not the goal of this effort to create a “secure world.” Rather, knowing that the world will never be completely safe, the cyber world should, in the context in which it is being used (e.g., social networking, financial transactions), be reasonably safe and, in some cases, safer than the physical world. Or, put another way, in most routine daily affairs, people should be able to make trust decisions that turn out to be right. In addition, the number of trust decisions people are asked to make should be limited (machines should apply our preferences for us when possible) and, when users or system administrators are asked to make trust decisions, they should be presented with meaningful information and an intuitive interface that encourages the right choices.

[8] It is true that the Internet poses some unique challenges in this regard. For example, although home security may be weak, there are proactive deterrents (e.g., neighborhood watches, police patrols) and reactive deterrents (e.g., arrest and prosecution), and the threat has natural limits (a limited number of skilled burglars and physical limits regarding how many homes can be burglarized in a day). By contrast, there is currently little in the way of cybercrime deterrence; attacks can be scripted, thus making anyone an “expert” hacker; and the amount of data that can be stolen is limited only by bandwidth.

VII. The Path Forward

There are essentially five major security components required to help facilitate trust, whether the "thing" being trusted is a person, device, operating system, software application, or piece of data. In the discussion that follows, we only describe identifiers, authentication, authorization, access control, and audit processes or services; we are not prescribing particular policies or mechanisms (e.g., enrollment mechanisms).

1. Identity Claims. Who does the person or what does the device or software claim to be? As a starting point, someone may claim to be a given person (e.g., John Smith) or simply claim to have a certain attribute (e.g., I am over 18 years of age). A device may claim to be an eBay server or a router, and an application may claim to be a particular version of Microsoft Office Word. The claim may also relate to source or integrity (this is a packet from an X Company router, or this spreadsheet was sent from John and has not been altered since being sent).

   An identity claim is, of course, only one part of the equation; in many contexts, reputation is equally critical and (especially because it is hard to speak about identity in absolute terms) will serve to add additional layers of assurance to an identity claim. This will be the case regardless of which element in the stack the claim attempts to validate. Robust reputation policies, processes, and systems will need to be built out to support the many trust decisions people need to make. Put another way, if a person claims to be John Smith, but you have never met John Smith before, the identification does not provide enough information to warrant a trust decision. Thus, closely related to the issue of identity are other attributes that are linked to that identity (e.g., past experiences, relationships, reputation).

2. Authentication. We must have mechanisms that allow identity claims to be verified. In the physical world, we often turn to formal documents (John Smith may have a national identity card, a passport, or a driver's license) to verify identity, even if the item used was not created for that purpose (e.g., a driver's license may be used by a bartender to ensure someone is old enough to purchase alcohol even though the intended purpose is to prove the right to operate a vehicle). We also have people whose function it is to verify identity (e.g., the notary public for documents, the Post Office for passport applications). There are clearly electronic analogies; we may use certificates to identify a device, or digital signatures to identity the author of software, and a root certificate for the organization verifying that claim.

3. Authorization policies. Assuming an identity is authenticated, there is some formal or informal policy that permits or prohibits activity based upon that authenticated identifier. Also of importance is who gets to determine the policy.

4. Access control mechanisms. Consistent with policy, a person may request access to a resource (e.g., the liquor store in the physical world, or an e-mail account in an electronic world). Access will be granted or denied based upon policy and verification of any necessary attributes. At times, people may obtain access to resources without, or in excess of, authority, thus potentially violating computer crime laws.

5. Audit. All the above (identity claim, proof of authentication, policies for authorization, request for access, decision on the request, and any unauthorized access attempts) must be documentable, as opposed to documented. How much audit data is collected, retained,

analyzed, and disseminated will depend on numerous factors, including the level of security required, the cost of collecting and storing audit data, and regulatory requirements.

Before describing the trusted stack in greater detail, it is important to note that this construct is not new[9] and work has been done in all these areas, albeit not with equal emphasis, for some time. For example, for the past 20 years the computer security industry has focused heavily on building stronger, more complex user authentication systems, resulting in improved password-based systems, two- or multi-factor authentication systems utilizing an array of tokens and biometrics, and more secure and better-performing implementations of Kerberos. Unfortunately, absent strong proof of identity or an identity claim, authorization is often granted to someone or something inappropriately, and what happens next cannot be traced back to its true source. Absent management tools, users still struggle with access control list maintenance. And absent audit, it is often difficult for system administrators to know whether their systems have been compromised, or how to quickly and reliably reconstruct events and conduct meaningful damage assessments, even if the source of the attack remains unknown. In short, too many key pieces remain inadequately developed and the pieces are not integrated sufficiently.

A. Trusted Devices[10]

Because all software operates in an environment defined by hardware, it is critical to root trust in hardware. Today, many computers come with a Trusted Platform Module (TPM),[11] a technology that will expand and enter new form factors. Moving security foundations to the hardware permits many beneficial scenarios. For example, some organizations have spent considerable resources combating hackers who have accessed their systems and engaged in interactive sessions. If machines did a machine-to-machine-based authentication rooted in TPM keys before allowing a network connection, then one could arguably exclude unapproved machines from accessing network resources. Using new cryptographic techniques, this can be done in privacy-compliant ways.

B. Trusted Operating System

The operating system must be verifiable based upon keys stored in the hardware (e.g., “trusted boot”). This allows the device to claim that the operating system has not been tampered with to bad effect.

Note that there are others things that must be done to increase trust in the operating system. Robust implementation of SD3 remains necessary since “trusted boot” does not by itself mean the operating system will be free from unintentionally introduced vulnerabilities. In addition, and equally important to a small subset of customers, operating system development organizations must take steps to prevent

---

[9] See, for example, the National Academy of Engineering, “Secure Cyberspace” (“better approaches are needed to authenticate hardware, software, and data in computer systems and to verify user identities”). http://www.engineeringchallenges.org/cms/8996/9042.aspx.

[10] From a purist’s perspective, the term “trusted devices” may be an overstatement. See Defense Science Board Task Force on High Performance Microchip Supply (in discussing the offshoring of the microelectronics industry). The Task Force noted that “One unintended result of this otherwise sound industry change is the relocation of critical microelectronics manufacturing capabilities from the United States to countries with lower cost capital and operating environments. Trustworthiness and supply assurance for components used in critical military and infrastructure applications are casualties of this migration.” http://www.acq.osd.mil/dsb/reports/2005-02-HPMS_Report_Final.pdf. But the real question, in the context of this paper, is whether a user has sufficient trust to use his or her device for Internet activities, not whether the device is completely trustworthy.

[11] TPM is a microcontroller that stores keys, passwords, and digital certificates. For more information on TPMs, see https://www.trustedcomputinggroup.org/faq/TPMFAQ/.

insertion of malicious code by members of the development community. To the extent that End to End Trust has been realized, robust authentication may limit opportunities for the insertion of malware by restricting access to code bases, and auditing of internal business activities should permit the provenance of bad code to be determined, thus allowing for a more robust response process.[12]

Finally, it is important to have a manifest of all the pieces that should be installed and not rely just upon mechanisms such as code signing for this purpose. If, for example, Microsoft issues a patch, an adversary can put the older version of the patched .dll on the machine (e.g., installation through social engineering). Although the older .dll will pass a signature test, the machine has been deliberately regressed to reintroduce an old vulnerability. Although some scanning solutions exist, there is no comprehensive way of knowing—or telling users—that they have the right components installed, a fact that can implicate not only security but also reliability.

C. Trusted Applications

Computers were, of course, designed to run code, without concern about its authorship or the intent of that author. Today there are multiple ways to help protect people from software vulnerabilities and malicious code. To protect users from vulnerabilities, code can be rewritten in safer languages, checked with analytic tools, compiled with compilers that reduce vulnerabilities (e.g., buffer overruns), and sandboxed when executed. To protect against malicious code, there are firewalls, anti-virus programs, and anti-spyware programs.

But although these approaches make users safer, criminals are not deterred by such preventive measures. To increase accountability, there is another effort that must be undertaken: code signing so that source can be better identified. Knowing source permits users to consider prior experiences, reputation, and other factors in deciding whether to install software. This is, of course, more problematic than it sounds for a host of reasons. For example, many exploits use code injection to bypass the loader, which checks to make sure code is signed.

Assuming users routinely reject unsigned code, the market response will be to provide signed code. Even if code is signed, however, it will still fall into one of three buckets. There will be code that is signed by a known entity (e.g., Microsoft, Oracle, Adobe) that is trusted due to past experience, brand reputation or some other factor; there will be code that is signed but known to be malware (e.g., spyware, which can then be blocked); and there will be code signed by entities that are not known to the user. Depending upon the criteria for obtaining a signature, the signature process itself may provide some deterrent to misconduct, much as extended validation certificates do today by providing a more extensive background investigation of the organization seeking the certificate. If code-signing signatures remain easy to obtain with no proof of physical identity, then any deterrent effect is lost and users have no assurance that malfeasance caused by the code can be addressed.

Even assuming the signing process is robust, users may not find signing sufficient to make a trust decision. Although users could address such concerns by simply refusing to run any code from a source not very well known, this would seriously undermine some of the advantages of the software economy: low barriers to entry and inexpensive global distribution channels. Thus, to support the growth of the software market, a reputation platform will also be needed to provide users with data about software publishers. This data may come from many sources: expert reviewers and

[12] Most governments accept that commercial software development practices cannot guarantee code purity.

researchers, other users, and reports of complaints (e.g., to consumer organizations, business organizations, and governments). Finally, it must be recognized that for a host of reasons, users will occasionally make the wrong "trust" decisions and that some code may not be signed. Thus, it is critically important that we create sandboxes to limit the damage that malicious applications can cause and that, where possible, we give users the ability to roll back transactions (which is another way to remediate harms). Similarly, it is important that the policies governing the reputation system include remediation processes to reverse or alter negative reputation decisions made based on inaccurate or incomplete data.

D. Trusted People

Perhaps one of the most famous Internet jokes was an early one. The New Yorker magazine ran a picture of two dogs at a computer screen with the caption, "On the Internet, nobody knows you're a dog."[13] That remains in large part true, even as child safety issues have driven people to look—mostly unsuccessfully—for ways to distinguish minors from adults.

The problem relates to how identity is determined in an electronic world. It is well-known that there are three ways to establish identity: what you know (e.g., a shared secret), what you have (e.g., a token or smartcard), and what you are (a biometric). For the most part, electronic identities have been established by having people disclose information that is known only by parties to the transaction, information sometimes called a "shared secret" (for example, your mother's maiden name). In the Internet context, this form of enrollment is no longer a sound method. The problem is that these shared secrets are increasingly stored and accessible online and, due to the increasing effectiveness of search tools and the increasing number of data breaches, shared secrets are no longer secret at all. In sum, the claim of identity is not robust and the authentication mechanism is flawed.

A safer Internet needs to support the option of identities based directly or derivatively upon in-person proofing, thus enabling the issuance of credentials that do not depend upon the possession of a shared secret by the person whose identity or identity claim is being verified. To some extent, government activities and markets themselves are driving in-person-proofing regimes. For example, many governments are issuing (or considering issuing) e-ID cards for government functions. But, to be clear, in-person proofing need not be controlled by governmental or quasi-governmental organizations; banks often have relationships with their customers that start with branch visits, schools have relationships with students and may routinely take in-person attendance, and employers know their employees and often issue identity cards based upon in-person proofing.

The creation of a distributed identity system that avoids shared secrets and has in-person proofing at its base has another salutary purpose: it allows us to devalue personally identifiable information (PII) and make a serious effort to reduce identity theft. As the ID theft problem has grown, most have focused on educating consumers not to disclose PII improvidently and ensuring that organizations that hold PII take reasonable steps to protect it.[14] Although these are important steps in terms of limiting identity theft, this approach will not reduce the occurrence of identity theft to the degree necessary. Although some consumers may make good judgments about disclosing PII, and some companies may engage in reasonable security practices and avoid data breaches, identity theft will

[13] The New Yorker, Vol.69 (LXIX) no. 20, July 5, 1993.

[14] In *In the Matter of BJ's Wholesale Club, Inc.,* the FTC held BJ's Wholesale Club liable for failing to protect personal data, finding a lack of reasonable security practices. It was the first time the FTC found a violation of Section 5 of the FTC act without any finding that the company misrepresented its security practices.

remain rampant so long as some large number of consumers and some number of data holders lose sensitive data.[15] This being true, it becomes clear that the key to combating ID theft is to devalue PII. If in-person proofing allows the issuance of true secrets (public-private key pairs), which can then be used for authentication, then criminals with access to PII do not have the key piece of data needed to consummate a transaction (e.g., obtain a line of credit at a bank), and the value of both social engineering attacks and intrusions into databases containing PII drops.

E. Trusted Data

As noted previously, one can identify the source of data and know whether the data has been altered without authority after being signed. Applications should incorporate seamless mechanisms for applying signatures to their outputs, and read signatures before opening documents, so that data origin and data integrity can be easily checked.[16] At the same time, management tools should permit users to apply policies based upon data origin and integrity so that fewer ad-hoc trust decisions are required.

While it may be important to know the source of data, it is also important to ensure that data is not accessed by unintended recipients. One of the benefits of creating this authenticated infrastructure is that it also permits senders to restrict access to data to authenticated individuals. This is an important privacy protection; far too often, sensitive data is shared too broadly or is too easily accessed by unauthorized individuals. As the firewall continues to diminish in importance, it is important to focus on protecting data as opposed to simply protecting the machines that store such data. Using the trusted stack to limit the flow of data mitigates the privacy harms that stem from unauthorized data flows and unauthorized data access.

Finally, as with applications, sandboxing should permit data to be opened in logically confined domains so that harms resulting from the combination of malicious content and the inevitable (even if diminishing) residual vulnerabilities can be contained.

F. Audit

An audit trail is a record of a sequence of events from which a history may be reconstructed. An audit log is a set of data collected over a period of time for a specific component. A series of audit logs can be studied to determine a pattern of system usage that, over time, can be used to highlight aberrant behavior such as criminal activity or the existence of malware.[17] Audit data is also necessary to roll back suspicious or harmful transactions. The ability to collect, report, and safely store audit data is an element of any compliance effort but, unfortunately, there is at present no ecosystem audit strategy that permits the rapid sharing of audit findings to create a common operational understanding among multiple parties. For example, policy languages and security audit logs are not standardized, making it difficult to correlate data sets across computational threads. Moreover, a lack of tools makes it difficult to analyze the data collected and turn that data into usable information. Finally, and notwithstanding the importance of audit in managing systems and ensuring accountability, there are

---

[15] Recent events suggest that data breaches will continue to occur and be serious. See http://privacyrights.org (as of Feb. 25, 2008, over 218 million data records of U.S. residents had been exposed due to security breaches since January 2005. Other nations' citizens have fared even worse. See http://www.computerworlduk.com/management/government-law/public-sector/news/index.cfm?newsid=6298 (U.K. government loses 25 million child benefit records).

[16] Similar mechanisms could be used to encrypt the content of data, thus enhancing communications privacy.

[17] Auditing can also be used, of course, to measure component performance, a topic beyond the scope of this paper.

no thriving standards bodies focused on standardizing audit data formats and tools. This is not to say, of course, that organizations do not collect audit data today. To the contrary, some complain that they are collecting too much data and what they really need are tools to better analyze the data that they already collect. But collecting the right type and amount of data and developing better tools to analyze that data are not mutually exclusive efforts. In fact, they help each other.

So what would the optimal audit system look like? It would begin with standard product and service instrumentation, meaning that code would be instrumented appropriately to collect a base level of audit data with management interfaces that would allow the amount of data collected to be increased or decreased depending on policy and current events (e.g., during an attack, the amount of data collected and analyzed might be increased). As is currently the case, audit data would be maintained in a distributed fashion by the individual, enterprise, Internet service provider (ISP), or government agency that manages a given resource, and each entity would decide to what extent this log data would be distributed or centralized within its own organization. The sharing of this information between organizations could be governed by policy and/or regulation. For example, government access might be governed by privacy laws, and private sector collection and sharing might be based upon existing fair information principles or regulatory requirements. In the end, this audit data would help organizations prove that they have fulfilled their obligations (i.e., compliance with business rules and regulations) and would provide a more robust way to catch bad actors on the Internet.

The primary challenges in attaining a more robust audit system are not all technical; there is an immense need for coordination across the industry to build some of the foundational pieces. Product teams must coordinate instrumentation, know where in the code audit calls should be made, and ensure the appropriate collection of audit data, all the time recognizing that the goal of an audit might be simple (e.g., determine whether a known individual has attempted to access a particular resource) or complex (e.g., determine whether multiple network scans from different IP ranges are actually emanating from the same source and occurring as a result of a single attack). Today, some products provide little auditing and others collect different audit data in different formats, making correlation difficult. There are also few analytic tools that can analyze audit data created in widely distributed, diverse product environments. Therefore, international standards are needed to aid the collection and analysis of audit data across heterogeneous systems, and management and analytic tools are needed to turn collected data into usable information.

In light of the security challenges faced—most notably, the international nature of most attacks—early work should focus on that subset of data needed to increase accountability in global networks. In most cases of unauthorized network use, a small subset of data would be needed to find source including connection date, connection time, port, protocol, and identifiers that link back to the source of an attack. Although some of this data already exists (e.g., systems do keep logs of connections and source IP addresses), it is not passed between the links in the chain, thus creating a laborious manual effort in an attempt to trace it back to the source. This is the point of the audit identifiers or tags; the goal is not to track all data flows, just those that are interesting and for which some action can be taken, whether it is finding the original source of a communication or permitting a router to drop dangerous packets.

Even assuming audit data were collected and could be passed when appropriate, other issues remain. Although cost of data storage continues to decline, there will remain questions regarding whether this data will be routinely collected and stored or whether it will only be collected when a

triggering event occurs. The latter is, of course, cheaper and more sympathetic to privacy concerns; the downside is that once events are detected it would not be possible to identify easily when the attacks started and how successful they were before being recognized. This is why it is critical to have standardized code instrumentation and management tools that are capable of collecting variable amounts of audit data as warranted by a user's own risk profile.

VIII. The Obvious Challenges

Clearly, there are obvious challenges when one even suggests an authenticated and audited infrastructure. Those challenges may be social (and political, as politicians are engaged in protecting social values through laws), economic and technical. The goal here is not to describe every challenge, but some of the core challenges. During this discussion, it must be remembered that social mores regarding privacy, free speech, and other values differ from country to country and, sometimes, even within countries. Although compatibility between different regimes is important in an increasingly global society—and contradictory laws may make it difficult for individuals to understand their rights and for businesses to remain compliant with national laws—complete harmonization is not achievable for a host of reasons. Thus, it is important that technology provide solutions that are adaptable to various social standards and regulatory environments.

With that background in mind, it remains true that ensuring that people can be identified raises the most complex social, political, and economic issues, with the No. 1 issue being privacy. The concern is twofold: (1) if authenticated identity is required to engage in Internet activity, anonymity and the benefits that anonymity provides, will be reduced; and (2) authenticated identifiers may be aggregated and analyzed, thus facilitating profiling.

Although anonymity may exist on the Internet due to historical evolution, the fact is that it serves many useful purposes. The fact that anyone can connect to the Internet without paying for the costs of an identification regime has certainly enhanced its growth. Moreover, as the Internet is imbued with free-speech characteristics, anonymity supports important policies regarding the promotion of free speech, even if harm sometimes occurs because of the anonymous nature of the communication. Indeed, it is important to remember that some societies have long accepted and promoted anonymous speech even if harm may occur. It remains possible to make anonymous phone calls (pay phones being replaced with disposable cell phones), and one can mail anthrax with no return address. There are both practical and philosophical reasons to continue to permit anonymous speech notwithstanding the risk of harm.[18] That said, it is an overstatement to say that the Internet, although imbued with speech, is simply about speech, or that the Internet is akin to other forms of communications networks. For example, the Internet provides communications abilities of a scale previously unknown; yes, one can call or send mail to millions of victims, but the time and cost makes this infeasible.

[18] It might be impractical, for example, to require all people to show a government ID before using postal services. From a philosophical point of view, anonymity promotes political speech. See *McIntyre v. Ohio Elections Comm'n*, 115 S. Ct. 1511 (1995) (in striking down an ordinance that required the disclosure of personal identity on political leaflets, the court noted that the ban on anonymous speech is not justified by the state's asserted interest in preventing the distribution of fraudulent and libelous information).

Of course, the multipurpose nature and power of the Internet make a comparison between the Internet and other communications technologies dangerous. Comparisons to traditional voice networks fail because the Internet is not just about—or even primarily about—voice. Comparisons to television fail because there is no scarcity of bandwidth, the primary underlying justification for content regulation. Comparisons to highways fail because people expect a greater degree of government involvement in public activities (e.g., driver licensing, car registration, insurance requirements) than in activities that are, in part, private. Notwithstanding the moniker "information superhighway," many people access the Internet from the most private of places—their home—where government involvement is unwelcome. As a result, changes in the Internet identity model may be worrisome to some.

Authentication also raises concerns about unique identifiers and profiling. Four things are important, here: (1) that many forms of identity will exist to allow users to provide different identifiers in different contexts, thus reducing the risk of profiling; (2) that users remain in control of what information they pass and when;[19] (3) that social rules support anonymity in appropriate contexts; and (4) the principle of data minimalization applies, so that identifying information is not collected without adequate justification.[20]

Clearly, this approach will not satisfy those who see the Internet's anonymity as the ultimate protector of privacy. This may particularly be true in those cases where anonymity promotes and protects unpopular speech. But the fact remains that if we hope to reduce crime and protect privacy, we need to give users the ability to know with whom they are dealing (if they so choose) and law enforcement the capability to find bad actors. It is also important to remember that there are multiple privacy interests at stake here; for example, in the e-mail context it is not just the sender of a communication who may have a privacy interest, but the recipient may wish to be left alone. Indeed, any regime should not only seek to provide greater authentication to those that want to provide it or consume it, but also provide anonymity for those who wish to engage in anonymous activities. Users should be able to choose to send anonymous communications, and users should be able to choose to receive mail only from known sources. Users who want to accept anonymous communications should be free to do so, but they should also understand that they may have little recourse if that anonymous communication proves to be harmful (e.g., a threat, a scam). The bigger "philosophical" issue relates to the fear that if an authenticated infrastructure is available, then neither market nor social forces will support a vibrant anonymous culture. Put another way, if authentication were possible, what if every social networking site, e-mail system, and Web site required authenticated identities? How would the social values promoted by anonymity be supported?

Although this debate cannot be resolved to everyone's satisfaction since it is impossible to prove, *a priori,* what will happen, one could argue that (1) people have long shown an interest in and support for anonymity; (2) markets will support anonymity, much as one can shop today without providing proof of identity; and (3) anonymity could be ensured by regulation, at least in some contexts. That said, some will still fear that the social changes upon us (particularly in a world consumed by terrorism) may push anonymity to the fringes. Although this is unlikely to happen, one must return to

[19] For an important discussion of these issues, see Kim Cameron's Laws of Identity, found at http://www.identityblog.com/stories/2004/12/09/thelaws.html.

[20] An example of this might be government legislation that requires authenticated identity when claiming government benefits or filing tax forms, but specifically prohibits requesting an authenticated identity when public information is being sought (e.g., health care information from a government Web site open to the public).

the fundamental point: An anonymous world cannot be the ultimate objective, either, particularly in a world marked by identity theft, attacks on critical infrastructures, and other events that require social response.

In addition to social forces, there are economic issues. Economic forces can drive certain behaviors, both good and bad, and those forces are often a function of decisions designed to stimulate economic activity and/or manage competing risks, even when those decisions do not serve security well.[21] For example, when Microsoft started down the path of extended validation certificates, the most obvious questions were, "Why should *new* trust criteria be established and who should ultimately be responsible for ensuring that only trustworthy organizations get certificates?" It remains true that in the United States, most e-commerce sites selling physical goods engage in domestic transactions due to shipping costs. And most companies doing e-commerce in the United States take payment by credit card. Since merchant banks do extensive investigations of businesses before authorizing them to take those credit cards, it would seem to be far more prudent to have the merchant bank issue the extended validation certificate, especially since the merchant bank's investigation is more extensive than the investigation conducted under the new standard.

But such a simple and elegant solution is thwarted by economic realities. Merchant banks cannot monetize extended validation certificates since consumers do not choose to shop exclusively at sites supporting such certificates, and there is no real economic incentive for e-commerce sites to pay for this service. Merchant banks may also worry about the liability that could be imposed for making an "assurance" statement. Finally, since Internet transactions are "card not present transactions," it is the merchant, and not the banks, that suffers the fraud loss; there may be no economic incentive for the banks to "validate" the identity of shoppers and potentially move the economic risk from the merchant to the bank. Attacking the problem from the other direction—from the government side—may also be impractical. Although the United States Office of the Comptroller of the Currency did in fact require banks to move away from username or password authentication for online banking, it might be concerned that requiring merchant banks to issue extended validation certificates to their merchants would constitute an inappropriate interference in the marketplace.

What becomes clear from this discussion is that a mix of social, political, and economic issues may need to be considered in combination when addressing the state of security on the Internet. For example, it was suggested earlier that ISPs could make dynamic trust decisions and use network access controls to protect the ecosystem. It is unclear, however, whether consumers (socially), regulators (politically), and access providers (economically) will accept scanning as a precondition to network access.

Finally, because governments have a primary role to play in investigating and prosecuting those who commit crimes, jurisdictional issues must be addressed. Much work has already been done in this area, from the G8's seminal work on computer crime, to the Council of Europe's Convention on Cybercrime. But these international agreements have their own limitations; the number of participating countries is limited, while the Internet is truly global. It remains true that applying sovereign laws to a sovereign-agnostic Internet is challenging; thus, responding to computer crime,

[21] One example of this relates to credit card usage on the Internet. To encourage users to shop online, the credit card associations agreed to waive the $50 liability borne by card users when their cards are used unlawfully. Although waiving the limit did encourage users to shop online without fear, it also meant that shoppers need not worry about the legitimacy of the merchant since any loss would be borne by others.

which is often international in scope, is difficult. In some cases, criminals may hop through countries to make it harder to collect evidence and identify source; in other cases, victims may be dispersed throughout the world and each case, standing alone, may not reach the economic thresholds necessary to warrant intense law enforcement attention.

IX. Conclusion

Much good work has been done to improve the security and privacy of computer users. But a key question remains: As we become increasingly dependent on the Internet for all our daily activities, can we maintain a globally connected, anonymous, untraceable Internet and be dependent on devices that run arbitrary code of unknown provenance? If the answer to that is “no,” then we need to create a more authenticated and audited Internet environment—one in which people have the information they need to make good trust choices.

It is also critical to understand the end goal: a more secure and trustworthy Internet ecosystem. In addition to empowering users to make good trust choices, the more general goals are to (1) mitigate common risks, substantially, so that public faith in the safety of the IT ecosystem is restored and/or enhanced; (2) permit security professionals to reduce their current efforts to address existing threats and allow them to redeploy those resources to address more intractable risks; (3) make it more difficult to conjure up new criminal schemes because authentication and audit make it more difficult to complete crimes successfully; and (4) enable law enforcement to find and prosecute a greater number of cybercriminals, thus increasing deterrence on the Internet. To achieve these goals, it is important to address all of the complicated social, political, economic, and technical issues raised to ensure we end up with the Internet we want, one which empowers individuals and businesses, and at the same time protects the social values we cherish.

Appendix A

During the creation of this paper, many people were provided with drafts or heard briefings and provided extremely helpful comments. In some cases, some individuals provided cumulative comments from their teams and I do not have a complete list of reviewers. In other cases, I presented this concept in this paper at organized events and received helpful comments in hallway conversations after the event. I apologize, in advance, for failing to recognize everyone individually.

First, I want to thank Steve Lipner, Ellen McDermott, and Phil Reitinger for reviewing numerous drafts of this paper without tiring or complaining.

Second, I want to thank the many people within, or affiliated, with Microsoft who provided thoughtful comments and ideas that substantially improved this paper. The list includes, but is not limited to, Pat Arnold, Marc Berejka, Eric Bidstrup, Bill Billings, Christopher Budd, Doug Cavit, Kim Cameron, Jerry Cochran, Peter Cullen, Jules Cohen, Chuck Cosson, Jon DeVaan, Pierre de Vries, Sean Finnegan, Jeffrey Friedberg, Tom Gemmell, Cristin Goodwin, Adrienne Hall, Todd Kutzke, Butler Lampson, Douglas Leland, Brendon Lynch, John Michener, Kevin Murphy, Mark Miller, Craig Mundie, Paul Nicholas, Rob Sonderman, Adam Shostack, George Stathakopoulos, Matt Thomlinson, and Cyril Voison. I also want to thank the Chief Security Advisors and Strategic Security Advisors.

Finally, I want to thank those outside contributors who reviewed this paper and provided critical feedback. This includes the Trustworthy Computing Academic Advisory Board (its members are identified at http://www.microsoft.com/presspass/press/2008/feb08/02-26TWCAABPR.mspx) as well as various individuals who saw presentations of the concepts in this paper and were kind enough to provide feedback after the event.