

A graphic consisting of a central yellow point from which several lines radiate outwards, some in yellow and some in light blue, creating a starburst or network-like effect.

**TwC Next**  
Marking a Milestone.  
Continuing Our Commitment.

**Microsoft** | Trustworthy Computing

# Trustworthy Computing Next

**Scott Charney\***  
**Corporate Vice President**  
**Trustworthy Computing**  
**Microsoft Corporation**

February 28, 2012

Version 1.01

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

# Trustworthy Computing Next

## Contents

I.	Trustworthy Computing – An Introduction .....	3
II.	The World Changes, Yet Again .....	4
A.	Living in a Data-Centric World .....	4
B.	The Role of Governments .....	8
C.	The TwC Pillars.....	9
III.	Security .....	11
A.	Background.....	11
B.	The Evolving Threat Model – Living With Persistent and Determined Adversaries .....	14
C.	The Cloud and Big Data .....	15
IV.	Privacy.....	17
A.	Background.....	17
B.	The Cloud and Big Data .....	18
C.	Government Access to Data .....	23
1.	Domestic Investigations.....	24
2.	International Investigations.....	25
V.	Reliability .....	29
A.	Background.....	29
B.	The Cloud .....	31
VI.	Conclusion.....	34

## 1. Trustworthy Computing – An Introduction

On January 15, 2002, Bill Gates sent a memorandum to all Microsoft employees announcing the Trustworthy Computing (TwC) Initiative.<sup>1</sup> In that memorandum, he noted the importance of providing computing that was as “reliable and secure as electricity, water services and telephony,” and noted the key aspects of a trustworthy platform included availability,<sup>2</sup> security, and privacy. He also made clear that the initiative was not just about technology: “There are many changes Microsoft needs to make as a company to ensure and keep our customers’ trust at every level – from the way we develop software, to our support efforts, to our operational and business practices.”

A graphic representation of TwC with its four pillars and early workstreams:



<sup>1</sup> See Gates’ memo: <http://www.microsoft.com/about/twc/en/us/twcnext/default.aspx>. Following the memo, a whitepaper was published that fleshed out concepts and provided more granular guidance as to how the company should proceed. See Trustworthy Computing whitepaper by Mundie, de Vries, Haynes and Corwine, <http://www.microsoft.com/about/twc/en/us/twcnext/default.aspx>.

<sup>2</sup> The memo also used the term reliability and that term was ultimately adopted as one of the four pillars.

The pressing need for Trustworthy Computing stemmed from the evolving role of computing in society. As Bill noted, “Computing is already an important part of many people’s lives. Within ten years, it will be an integral and indispensable part of almost everything we do.” This last statement could not be truer. Over the last ten years, we have witnessed the rise of the Internet citizen with members of society connected through email, instant messaging, video-calling, social networking, social searching, and a host of web-based and device centric applications. We now live in a world where humans are more connected by and reliant on computing technology than ever before.

While this new world creates great opportunities, we are faced with both old and new challenges. While the security of the power grid, the global financial system, and other critical infrastructures have long been of concern,<sup>3</sup> new threat models involving persistent and determined adversaries and the specter of cyber warfare have raised new challenges for computer security professionals. The proliferation of connected devices and a massive increase in the amount and types of data available for collection, analysis and dissemination have strained traditional rules to protect privacy. And with people dependent on devices, cloud services, and anytime/anywhere access to their data, the reliability of information systems has taken on greater importance. In sum, a relentless focus on Trustworthy Computing has never been more important. So now, ten years later, how should Trustworthy Computing continue to evolve?

## I. The World Changes, Yet Again

Although Trustworthy Computing was launched with an appreciation of the new role computers were playing in our lives, it is important to understand how the world is changing yet again and how those changes affect the four pillars of TwC. The two most profound changes relate to data-centricity and how governments are engaging in Internet related issues.

### A. Living in a Data-Centric World

We are moving to a world of connected devices (including phones, computers, televisions, cars, and sensors) with devices now outnumbering the humans they serve.<sup>4</sup> In this world, there are old forms of data (e.g., bank records, telephone records), but also new forms of data that may be particularly revealing (e.g., user

---

<sup>3</sup> See, for example, the 1997 President’s Commission on Critical Infrastructure Protection ([http://itlaw.wikia.com/wiki/Critical\\_Foundations:\\_Protecting\\_America%E2%80%99s\\_Infrastructures](http://itlaw.wikia.com/wiki/Critical_Foundations:_Protecting_America%E2%80%99s_Infrastructures)) and the 2008 CSIS Commission on Cybersecurity for the 44th Presidency (<http://csis.org/program/commission-cybersecurity-44th-presidency>).

<sup>4</sup> It was expected that the number of Internet devices would reach 5 billion in August 2010. <http://www.cellular-news.com/story/44853.php>. Cisco has also predicted that by 2015, the number of Internet connected devices would be twice the human population. <http://www.bbc.co.uk/news/technology-13613536>.

created data on social networking sites and geo-location data). The separation between an individual's personal and professional life is eroding, as the "consumerization of information technology (IT)" – which refers to an individual's desire to pick IT devices and have those devices span one's professional and personal lives – conflates, or perhaps even obliterates, the line between two historically separate worlds managed by different entities with different rules. Perhaps most importantly, there are new centralized abilities to store, aggregate, search, analyze and disseminate this rich data. The ability now exists not only to build a historical trail of a person's activities, but to predict their future behaviors in new and interesting ways. It turns out, for example, that a reliable predictor of whether someone will default on their mortgage is not his or her credit score, but whether that individual's social networking friends have paid their debts.<sup>5</sup> This data richness also means we cannot only treat people who are ill, but potentially predict the diseases to which they are prone, a fact that could be used to save a life or deny health insurance.

Not all data is created equal, of course, and geolocation data deserves specific mention. It turns out that "location, location, location" is not just a mantra in real estate, but increasingly important in delivering real-time services to individuals. At the same time, however, it has been noted that "GPS monitoring generates a precise, comprehensive record of a person's public movements that reflects a wealth of detail about her familial, political, professional, religious, and sexual associations....Disclosed in [GPS] data . . . will be trips the indisputably private nature of which takes little imagination to conjure: trips to the psychiatrist, the plastic surgeon, the abortion clinic, the AIDS treatment center, the strip club, the criminal defense attorney, the by-the-hour motel, the union meeting, the mosque, synagogue or church, the gay bar and on and on").<sup>6</sup>

The changes occasioned by this new world are daunting; there are obvious and non-obvious risks. To understand these issues more fully, it is important to appreciate how the information technology model is changing. When the term "World Wide Web" was coined in 1990, it referred to a web of documents viewed by users in a client-server architecture.<sup>7</sup> In this world, the threats were linear.

---

<sup>5</sup> See "As Banks Start Nosing Around Facebook and Twitter, the Wrong Friends Might Just Sink Your Credit," <http://www.betabeat.com/2011/12/13/as-banks-start-nosing-around-facebook-and-twitter-the-wrong-friends-might-just-sink-your-credit/>.

<sup>6</sup> See *United States v. Jones*, <http://www.supremecourt.gov/opinions/11pdf/10-1259.pdf> (Sotomayer, J., concurring, (citing *People v. Weaver*, 12 N. Y. 3d 433, 441–442, 909 N. E. 2d 1195, 1199 (2009)). It is worth noting that although the themes in this paper are universal, the text clearly has a U.S. focus. As former Vice Chair of the OECD Group of Experts on Security and Privacy, and former Chair of the G8 Subgroup on High-Tech Crime, the author believes it is impossible to be fully versed in the laws and cultures of every nation. As such, the author has written about that which is most familiar.

<sup>7</sup> [http://en.wikipedia.org/wiki/World\\_Wide\\_Web](http://en.wikipedia.org/wiki/World_Wide_Web).

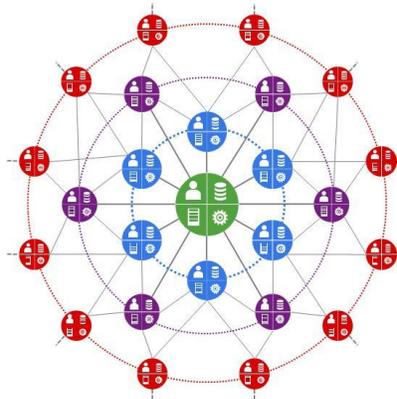
If one were to picture the original web, it might look something like this:



ICONKEY



Today, we still have a web, but it is not just of documents. Our experiences are far richer, in part because we have a web of users, machines, applications, and data. Additionally, while some of these connections are obvious (e.g., typing in a URL takes an individual to a web page), many are not (unknown to the user, that web page may in fact pull in data from other sites). This new world looks something like this:



ICON KEY



In such an environment, each and every element (user, machine, application, and data) may be helpful or harmful, innocuous or dangerous. In this world, the threats are no longer linear, but reflect the mesh of interconnectedness that has developed over the past decade.

In such an interconnected world, how should we think about the four pillars of trustworthy computing – security, privacy, reliability, and the business practices that reflect on those first three core attributes? How should we think about “units of manageability?” To some extent, it is about span of control. A user can be directly responsible for his or her own stack (e.g., patching a computer, downloading trusted applications, and backing up data). A company may run its own data center, hiring the system administrators, purchasing the hardware, leasing the software, and managing its data. In other cases, an entity may have a direct relationship with the party that is responsible for managing these elements. In that sense, a user can determine whether to trust the entity they are dealing with (e.g., by relying upon a company with a good reputation) and obligations can be set (e.g., through contracts, including terms of use, at least where the user has realistic alternatives and reasonable chance of gaining enforcement of terms.) Indeed, much of the current cloud adoption guidance falls into this tier. In yet other cases, the user is connected to elements which are far removed; elements over which the user may have no direct or derivative control to rely upon and, equally important, no visibility into important practices (or shifts in those practices, when they occur). How can trust be created in such remote relationships? How do we create a world where individuals have better transparency into remote business practices and can also ensure that the information they share with others is used in the way they intend?

One approach is to rely upon metadata and claims to help express information about appropriate data usage and the state of systems. Those responsible for managing technology could represent, in consistent ways, how they govern their business practices as well as the international standards with which they are compliant. Some of these international standards or accepted practices already exist. For example, there are ISO standards on security, Fair Information Practices for privacy; and standards regarding how to verify a person’s identity.<sup>8</sup> At times, compliance with those standards is revealed to end users; when two parties agree to support https://, a browser may present visible clues.

In the future, the presentment of attested to or verified claims – which can relate to the health of a machine, the identity of a person, and the provenance of software – will take on greater significance and, indeed, could be extended to other important attributes. For example, a software program could provide not just a signature to show provenance, but also indicate whether it was developed pursuant to a recognized secure development standard. Similarly, data may have metadata about its author, acceptable uses, and other important attributes (such as whether the data is personal or corporate or whether it contains personally identifiable information). Having such metadata should help enable the enforcement of rules relating to how the data can be used or shared. While such rules may be attached to the data itself, the rules could be stored separately from the data and the data could “point” to the applicable rule. This would allow someone who has

---

<sup>8</sup> NIST SP800-63, for example, provides different levels of assurance based upon the level of confidence in an asserted identity validity, which in turn stems from the proofing process used. If Level 1 involves self-assertion and Level 4 involves strong proof of identity, those relying on an identity assertion can require “Level 4” if necessary and, more importantly, know what that means. <http://csrc.nist.gov/publications/nistpubs/800-63/SP800-63V1.0.2.pdf>.

tagged data to later “change one’s mind” regarding how the data should be handled and enforce these changes by modifying the applicable rule. In sum, with users and computers connecting to and sharing data with elements which are far removed, creating a common mechanism for expressing and honoring policies may allow remote parties to make more meaningful trust decisions.

## B. The Role of Governments

This connected world of devices and services – and the creation of “big data”<sup>9</sup> – has been challenging for governments. It must be remembered that any transformative technological change which recasts the way we live our lives – and enables potentially troubling criminal and nation-state activity – will engender government responses. But a government’s “relationship” with the Internet is actually a complex one, as governments are users (they are, essentially, large enterprises with citizens as customers), protectors (of both the rights of individuals users as well as the Internet itself), and exploiters (there has long been military espionage on the Internet<sup>10</sup>).

As users, governments are concerned about the same issues as other Internet users: how can we protect the security and privacy of the data throughout its lifecycle, across systems and devices? How can we ensure that my system is available when needed, particularly in times of crisis? Should our agency embrace the cloud and, if so, how do we ensure our requirements are met by the cloud provider? These are questions for all Internet/cloud users, although governments may face a unique set of adversaries and unique accountabilities (e.g., to voters as opposed to customers).

Governments have also become increasingly attentive to protecting the rights of Internet users (e.g., privacy, security, freedom of speech and association) as well as protecting the Internet itself. While governments can leverage market forces through their purchasing power, market forces standing alone are designed to respond to market demands, not necessarily meet the unique requirements sometimes occasioned by public safety and national security concerns. Simply put, it is hard to make a market case for the Cold War. Governments are therefore looking at the other tools at their disposal, such as passing regulations to protect security, privacy, and/or reliability, or using law enforcement and military responses to more effectively deter harmful conduct. Such law enforcement and military responses are hampered by the difficulty in reliably and definitively tracing an attack back to its source.

Unlike traditional attacks involving significant military aircraft or troop movements, Internet attacks, even if against traditional military targets (e.g., a computer system operated by the military in times of war), may be

---

<sup>9</sup> “Big Data” refers to very large data sets. It is believed that big data will be the next frontier for innovation, competition, and productivity. See [http://www.mckinsey.com/Insights/MGI/Research/Technology\\_and\\_Innovation/Big\\_data\\_The\\_next\\_frontier\\_for\\_innovation](http://www.mckinsey.com/Insights/MGI/Research/Technology_and_Innovation/Big_data_The_next_frontier_for_innovation).

<sup>10</sup> See Cliff Stoll, *The Cuckoo’s Egg*, 1989.

launched by individuals without ties to a particular government (asymmetric attacks). Put another way, the myriad of Internet actors and motives, when combined with global connectivity and traceability challenges, makes an attacker's identity, location, affiliation and motive difficult to ascertain quickly and authoritatively. Moreover, because tracing the attack will often involve private sector parties, the need for a public/private partnership is clear but the rules for sharing critical information are not. Indeed, parties have good reasons *not* to collect and share data, including protecting the privacy of customers, avoiding potential legal liability, avoiding entanglement in government disputes, and avoiding the costs and potential risks of data retention.

There is continuing debate over whether the private sector should store data (some countries have mandatory data retention rules while others do not), give what data they have to governments, and whether governments should monitor private sector networks. In sum, a core challenge is how governments fulfill their traditional responsibilities without unnecessarily encumbering businesses, stifling important innovations, and inappropriately undermining civil liberties.<sup>11</sup>

Finally, while it would certainly be easier if governments were solely concerned with protecting the Internet, they have reasons to exploit it too. Economic and military intelligence may give a country a competitive advantage against friend and foe alike, and offensive military capabilities are seen as important in terms of both cyber conflict and as a force multiplier during kinetic warfare. It is not the point of this paper to discuss in detail the complexity of all of these roles and responsibilities, but rather to note that these issues, which are being explored deeply in other fora,<sup>12</sup> have caused governments to think in increasingly active and nuanced ways about the Internet and that, in turn, has implications for the future of the Internet.

### C. The TwC Pillars

In thinking about TwC Next, one obvious question was, "are the pillars described ten years ago still the right ones?" If one is focused on "trust" in computing, the answer is "yes." There are certainly many other important issues for society to address but, at their heart, these do not relate to "trust in computing." For example, it is important that devices be accessible to those with disabilities and that they be designed with environmental sustainability in mind. But while a device that cannot be used by all or wastes energy may be undesirable, that does not mean it cannot be trusted to perform its computing function in a secure, private and reliable way. Thus, while society must work on a wide range of issues – issues far broader than the four pillars of TwC – it remains true that these pillars have a unique role relative to trust in technology.

---

<sup>11</sup> Importantly, national differences affect the definition of "harm."

<sup>12</sup> See Owens, Dam and Lin, editors, "Technology, Policy, Law, and Ethics Regarding U.S. Acquisition and Use of Cyberattack Capabilities," [http://www.nap.edu/catalog.php?record\\_id=12651](http://www.nap.edu/catalog.php?record_id=12651), and Center for a New American Security, "America's Cyber Future: Security and Prosperity in the Information Age," <http://www.cnas.org/cyber>.

This is not to suggest that this new sensor laden, data-centric world does not alter our thinking about the pillars; indeed, it is interesting to think about both the inter-relationship and relative importance of these pillars in this new world. Historically – and perhaps still today – the most obvious intersection relates to security and privacy, where one can find both harmony and conflict. To the extent security is a protector of privacy (one of the core objectives of security is to protect the confidentiality of data), there is harmony. Indeed, this explains why Fair Information Principles designed to protect privacy have long had a security principle.<sup>13</sup> On the other hand, security techniques can be privacy-invasive; consider, for example, conducting traffic analysis and deep packet inspection in an effort to find computer abuse. Finally, there may be times when security and privacy have little intersection. Updating a system to patch a vulnerability may have little to do with privacy, just as using data in a way not described in a privacy notice may have little to do with whether that data was secured appropriately. What is important is that while security and privacy do sometimes overlap, it is not a “zero sum game” where security must be traded for privacy or vice versa. As we have noted in Microsoft’s efforts around online identity, the primary objective should be to promote both values at the same time, even if there are times when trade-offs must in fact be made.<sup>14</sup>

There are other interesting interplays among the pillars, even if not obvious. For example, reliability failures may be the product of security events (e.g., denial of service attacks) which may lead to increased network monitoring and, therefore, an impact on privacy. Additionally, to ensure the reliability of cloud services, service providers may distribute data geographically to ensure that a natural disaster in one part of the world does not singlehandedly cause the loss of data. However, placing data in another country may raise interesting security and privacy issues, in part because it may make the data available to foreign governments through local legal process or because the foreign jurisdiction’s laws may provide different levels of security and privacy protection for such data.

The cloud will also change the relative importance of the pillars. It is probably fair to say that, in the past, security was primary; privacy and reliability were secondary. Security is likely to remain the dominant concern in this new world, especially since security events have the capacity to impact privacy and reliability as well. That said, big data will exacerbate concerns about privacy and reliability failures may cause significant social and commercial disruptions. Thus, the relative weight of the pillars will become more equal and the world’s focus on these topics will have to be more evenly distributed. Additionally, concerns *within* a given pillar will become increasingly complicated due to the major shifts occurring in societal dependence on IT and complexities in the emerging IT model. These changes are discussed in more detail below when the individual pillars are discussed.

---

<sup>13</sup> See, e.g., the OECD Guidelines on the Protection of Privacy and Transborder Flows of Personal Data, [http://www.oecd.org/document/18/0,3343,en\\_2649\\_34255\\_1815186\\_1\\_1\\_1\\_1,00.html](http://www.oecd.org/document/18/0,3343,en_2649_34255_1815186_1_1_1_1,00.html) (Security Safeguards Principle).

<sup>14</sup> See “The Evolution of Online Identity,” IEEE Security & Privacy, vol. 7, no. 5, pp. 56-59, Sep./Oct. 2009.

While it is impractical to describe every social, commercial and technological implication of big data and government involvement, certain issues are rising to the forefront: the evolving threat model and its impact on the security of the Internet, the advent of big data and whether privacy will be lost in a potentially Orwellian age of data analytics, and what happens when society is increasingly (totally?) dependent on the availability of systems and data and those core assets become unavailable either due to normal events (e.g., hardware failures, software vulnerabilities, configuration errors) or deliberate attack. These issues, individually and collectively, drive new thinking that lies at the heart of TwC's future.

## II. Security

### A. Background

When Bill Gates announced Trustworthy Computing ten years ago, he focused on the four core pillars of security, privacy, reliability, and business practices. Notwithstanding this fact, most people initially equated TwC with security. This occurred partly because the announcement of TwC came after a series of important security-related events. In the summer of 2001, the Code Red worm had given Internet users an idea of the potential consequences of a malicious attack. Then, on September 11, 2001, a kinetic terrorist attack shut down the U.S. stock market for five days, in part because of the disruption to information systems. Recognizing that such a disruption to IT systems could have occurred through cyber attack, the events of 9/11 raised new concerns about critical infrastructure protection. Lastly, a week after the September 11th attack, the Nimda worm appeared, another virus with worldwide impact. Thus, the world focused its attention on security and Microsoft itself redirected considerable energy and resources to improving the security of its products. It did so by focusing on an approach called SD3: Secure by Design, Secure by Default, and Secure in Deployment. Secure by Design was about reducing vulnerabilities in code, a goal which the company pursued by mandating the use of the Security Development Lifecycle (SDL).<sup>15</sup> Secure by Default meant that products had fewer features turned on by default, thus reducing a product's attack surface. To ensure Secure in Deployment, the company updated its patching technology and processes, reducing the number of patch installers, building better patching tools and adopting a more predictable cadence for patching.

Microsoft also focused on non-technical solutions. It is important to appreciate that many of the online security threats facing customers do not exploit software vulnerabilities, but people's trusting nature. Our project called Broad Street, which measures how malware compromises computers, indicates that over 50

---

<sup>15</sup> The SDL involves building threat models at design time, and then architecting, building and testing products to help ensure those threats are mitigated. For a fuller description of the SDL, see <http://www.microsoft.com/security/sdl/default.aspx> and the book by Michael Howard and Steve Lipner, "The Security Development Lifecycle" (2006).

percent of compromises now involve social engineering.<sup>16</sup> Part of the problem is that making trust decisions is confusing and difficult – people rarely have the information they need to understand security risks and lack the guidance necessary to make good security decisions (the same can be said about privacy decisions too). There is an opportunity for the software industry to improve security (and privacy) by reducing the number of decisions that users need to make and by making the decisions that remain easier. To help achieve this aim, Microsoft has published guidance for designing trust user experiences, experiences that should be “NEAT”: Necessary, Explained, Actionable, and Tested. NECESSARY means that the user should only be involved in a trust decision if he or she has unique knowledge or context that informs the decision to be made. The steps for making a good decision should be EXPLAINED in clear terms. That explanation must be ACTIONABLE, in that users should be able to make good decisions in both benign and malicious scenarios. Finally, the experience should be TESTED, so it is shown to work for a wide range of potential users.<sup>17</sup>

Notwithstanding these technical and non-technical efforts, improving computer security remained challenging. This was due to both the nature of the target (the Internet provides global connectivity, significant traceability challenges, and rich targets), the various ways targets could be attacked (through supply chain attacks, vulnerabilities, system misconfigurations, and social engineering), and the adaptability of attackers (as operating systems became more secure, attacks moved up the stack to the application layer; as applications became more secure, cybercriminals turned their focus to social engineering).<sup>18</sup> These realities limited the effectiveness of Microsoft’s substantial early work on security and, recognizing this fact, Microsoft’s security strategy continued to evolve.

Among the more noticeable of these changes was our focus on establishing end-to-end trust, a strategy reflected in a 2008 white paper of the same name.<sup>19</sup> This updated strategy took a different and more granular approach to the problem of building security into IT systems. We recognized that SD3 was necessary but not sufficient, fundamental to success but not enough to achieve it. Thus, Microsoft started focusing more acutely on the entire IT stack, from the hardware to the software to the user. We also noted that solutions to difficult problems required that social, economic, political and IT interests and capabilities align; something that

---

<sup>16</sup> See [http://download.microsoft.com/download/0/3/3/0331766E-3FC4-44E5-B1CA-2BDEB58211B8/Microsoft\\_Security\\_Intelligence\\_Report\\_volume\\_11\\_Zeroing\\_in\\_on\\_Malware\\_Propagation\\_Methods\\_English.pdf](http://download.microsoft.com/download/0/3/3/0331766E-3FC4-44E5-B1CA-2BDEB58211B8/Microsoft_Security_Intelligence_Report_volume_11_Zeroing_in_on_Malware_Propagation_Methods_English.pdf).

<sup>17</sup> [http://blogs.msdn.com/cfs-file.ashx/\\_key/communityserver-components-postattachments/00-10-16-10-50/NEATandSPRUCEatMicrosoft\\_2D00\\_final.docx](http://blogs.msdn.com/cfs-file.ashx/_key/communityserver-components-postattachments/00-10-16-10-50/NEATandSPRUCEatMicrosoft_2D00_final.docx).

<sup>18</sup> In addition to the types of attacks, it remains true that security is very much an arms race between those attacking systems and those seeking to protect them. For example, once Microsoft created a predictable cadence for issuing patches on “Patch Tuesday,” hackers reversed engineered those patches and released malware on “Exploit Wednesday.” Microsoft responded with the Microsoft Active Protections Program (MAPP). In this program, MAPP partners receive vulnerability information early so that they can provide updated protections to customers via their own security mechanisms, such as updating virus signatures or network-based and host-based intrusion prevention systems. See <http://www.microsoft.com/security/msrc/collaboration/mapp.aspx>.

<sup>19</sup> See Establishing End to End Trust, [http://download.microsoft.com/download/7/2/3/723a663c-652a-47ef-a2f5-91842417cab6/Establishing\\_End\\_to\\_End\\_Trust.pdf](http://download.microsoft.com/download/7/2/3/723a663c-652a-47ef-a2f5-91842417cab6/Establishing_End_to_End_Trust.pdf).

frequently failed to occur. In addition to the End-to-End Trust work, we also focused thematically on several other key areas, including the changing threat model,<sup>20</sup> and how applying public health models to the Internet might proactively improve the state of security.<sup>21</sup>

Throughout this time, we tracked certain statistics to measure our success, which initially would have been defined as “reducing vulnerabilities in code” but later evolved to “making customers safer.” This shift is important, especially if one recognizes that it is impossible to reduce the number of vulnerabilities in complex products to zero. It is also important because it helps define one’s strategic direction: while we continued to improve the SDL to reduce vulnerabilities, it became increasingly important to focus on defense-in-depth.

The goal of a defense-in-depth strategy is to provide additional protections so that even products with vulnerabilities are harder to exploit successfully. Technologies designed to do that – such as DEP (Data Execution Prevention) and ASLR (Address Space Layout Randomization) – were integrated into the SDL and had their intended effect in reducing successful attacks. Statistics show that in the early years, there were significant drops in the number of Microsoft vulnerabilities, but ultimately our vulnerability reduction rates have leveled off as the number of vulnerability researchers increased and their tools got better. Notwithstanding this fact, computers running later versions of software that have security updates applied are less susceptible to infection by malicious software and most successful attacks are the product of social engineering and exploiting vulnerabilities for which fixes already exist.<sup>22</sup> More specifically, our Malicious Software Removal Tool (which cleans infected machines as part of the monthly automatic update process) shows that on 32 bit systems, Windows XP SP3 has 10.9 machines cleaned per 1,000 scans, Vista SP2 has 4.4 machines cleaned per 1,000 scans, and Windows 7 SP1 has 1.8 machines cleaned per 1,000 scans.<sup>23</sup>

What these statistics make clear is that investments in computer security do serve to blunt opportunistic threats, particularly when a computer user has conducted “basic computer security hygiene” such as deploying newer versions of products, patching vulnerabilities promptly, managing configurations carefully, and engaging in continuous monitoring (e.g., Anti-Virus, Intrusion Detection Systems). It is also clear that social engineering has proved a more intractable problem, since users remain prone to clicking on attachments and visiting dangerous websites, often in response to phishing.

---

<sup>20</sup> See Rethinking Cyber Threats, <http://www.microsoft.com/download/en/details.aspx?displaylang=en&id=747>.

<sup>21</sup> See Collective Defense: Applying Public Health Models to the Internet, available at <http://www.microsoft.com/mscorp/twc/endoendtrust/vision/internethealth.aspx>.

<sup>22</sup> See [www.microsoft.com/sir](http://www.microsoft.com/sir): [http://www.microsoft.com/security/sir/story/default.aspx#!0day\\_results](http://www.microsoft.com/security/sir/story/default.aspx#!0day_results) and [http://www.microsoft.com/security/sir/story/default.aspx#!0day\\_exploit](http://www.microsoft.com/security/sir/story/default.aspx#!0day_exploit).

<sup>23</sup> See SIR Report, pp. 73-80, [http://www.microsoft.com/security/sir/keyfindings/default.aspx#!section\\_4\\_2](http://www.microsoft.com/security/sir/keyfindings/default.aspx#!section_4_2)

## B. The Evolving Threat Model – Living With Persistent and Determined Adversaries

While the quality of code has improved and infections rates have declined, the threat model has continued to evolve in challenging ways. Leaving aside cyber criminals who have shown increasing sophistication, some organizations have also taken a far greater interest in computer security, including by honing their offensive skills. Opportunistic threats have been supplemented by attacks that are more persistent and, in many cases, far more worrisome. While these types of targeted attacks have been coined “Advanced Persistent Threats” or “APTs,” that term is a misnomer. While some of these attacks are “advanced” (as in “sophisticated”), many are not; rather, the attack vectors are often traditional and unsophisticated: unpatched vulnerabilities and misconfigurations (both of which can be exploited by simple, widely available tools), and social engineering. Whether advanced or not, what marks these attacks is that the adversary is persistent (willing to work over time) and determined (firmly resolved to penetrate a particular victim). Importantly, what has become clear is that if an organization is targeted with persistence by a determined adversary, a successful penetration or major disruption is likely.

The computer security community needs to adapt to this new world, one in which there are an increasing number of both opportunistic and targeted threats. This adaptation means embracing a two-pronged strategy. First, those managing IT systems must improve their basic hygiene to counter the opportunistic threats and make even persistent and determined adversaries work harder. This includes migrating to newer, more secure systems, patching vulnerabilities promptly, configuring systems properly (in part through increased automation), educating users about the risks of social engineering, and taking other steps – whether they involve people, process, or technology – to manage risks more effectively than done today. Enterprise security professionals may see this as nothing new, but it equally applies to home users that self-manage their IT systems as well as computing devices that are not actively managed at all.<sup>24</sup> This is another area where governments may play an increasingly active role, as some nations are looking at legislatively mandating the adoption of information risk management plans as well as greater disclosures of the risks facing those dependent on Information and Communications Technology (ICT) systems.<sup>25</sup>

---

<sup>24</sup> One approach is to apply public health models to the Internet in an effort to better ensure that home machines are healthy and infected machines are treated promptly. See *Applying Public Health Models to the Internet*, found at:

<http://www.microsoft.com/mscorp/twc/endtoendtrust/vision/internethealth.aspx>.

<sup>25</sup> An example of legislation includes The Cybersecurity Act of 2012, introduced in the U.S. Senate during the week of February 13, 2012. With regard to disclosure, the U.S. Securities and Exchange Commission recently issued guidance that “Registrants should disclose the risk of cyber incidents if these issues are among the most significant factors that make an investment in the company speculative or risky.”

See <http://www.sec.gov/divisions/corpfin/guidance/cfguidance-topic2.htm>.

The second part of the strategy involves fundamentally how computer security professionals address the persistent and determined adversary. In many of these cases, the attacks are marked by long-term efforts to penetrate a computer system stealthily and then leverage the fact that a hard perimeter, once defeated, reveals a soft interior that can be navigated easily for long periods of time. This being the case, the security strategy deployed for blunting opportunistic threats – a security strategy focused predominantly on prevention and secondarily on incident response – will not be enough. Instead, we must focus on four areas: prevention, detection, containment, and recovery.

While these elements are of course not new, there are opportunities to significantly increase our effectiveness in these areas. For example, while many organizations manage intrusion detection systems, security strategies have not focused on capturing, correlating and analyzing audit events from across the enterprise to detect anomalies that belie attacker movement. With big data unlocking new opportunities in a host of areas, we now need to explore how big data can create situational awareness in the security context while, of course, ensuring that potential privacy concerns are addressed. Additionally, notwithstanding how interconnected services have become, we need to focus on containment (e.g., network segmentation, limiting user access to least privilege) to ensure that, if part of a network is compromised, the adversary is well contained.

There is an important parallel here to past efforts to promote secure development. Many security professionals would argue that by 2002, the security industry had already developed security development techniques, tools, and in some cases, implemented defense-in-depth code protections. But such efforts were not undertaken at scale and some important lessons cannot be learned until efforts are made to formalize conceptual practices and apply them at scale, in decentralized environments, over multiple releases of a product. Like Microsoft's long-term commitment to secure development and the SDL, the techniques necessary for containment and recovery are not new but efforts to apply those techniques at scale and over time will undoubtedly provide equally valuable lessons in how to protect, detect, contain and recover from attacks. In short, old ideas must be supplemented by new efforts and these efforts must be undertaken at scale and with rigor.

### C. The Cloud and Big Data

In addition to the emergence of persistent and determined adversaries, the cloud and big data will also need to inform computer security efforts. At the highest conceptual level, Microsoft has frequently been asked: "is the cloud better or worse for security?" The answer is, somewhat unhelpfully, "yes." Put another way, there are aspects of the cloud that make it a more secure environment than the distributed IT environment we have today, but other cloud attributes make security more challenging. Most specifically, principles related to Secure by Design, Secure by Default, and Secure in Deployment must be augmented with a principle of Secure in

Operation. We begin with some broad observations about the move to the cloud, and then drill down more deeply on how to think about cloud security.<sup>26</sup>

At the highest level, the cloud offers at least four significant security advantages over the existing distributed systems model. First, the world has a significant shortage of computer security professionals, a problem that is not likely to be solved in the short term.<sup>27</sup> Indeed, many small organizations may have no IT staff, let alone security experts. To the extent cloud providers enable a centralization of security expertise in large data centers and adhere to emerging international standards of operational security, security may be increased dramatically over time.<sup>28</sup> Second, the centralization of data, combined with better monitoring tools, may permit better protection than exists in today's massively distributed world, where monitoring may not be rigorous in all organizations and correlating data among organizations remains difficult. Third, much of the code created for the cloud itself (e.g., Windows Azure, an operating system for the cloud, and virtualization technologies) was created after the implementation of secure development practices like the SDL, thus helping to ensure better security code quality. Fourth, in a world increasingly composed of sandboxed applications that are gated through "app stores," more users may be downloading applications from known sources (or sources for which there is substantial reputational data) and those sources may have better security controls in place to respond if an application proves malicious.

At the same time, the emergence of the cloud will no doubt result in new threats, some of which are discussed actively today and others which will require deeper analysis. Security professionals recognize, for example, that the cloud will lead to massive data consolidation and, relatedly, present a rich target for cybercriminals. Additionally, access to that rich data set is too often gated by authentication mechanisms that remain far too dependent on usernames and shared secrets (passwords) that are too easy to steal, divine or solicit, either through social engineering the end users themselves or a call center that can mistakenly grant access to an unauthorized individual. This being true, we continue to need a much stronger identity metasystem, one that makes it considerably harder to spoof another individual.<sup>29</sup> Third, there will be criminals who try to exploit the internals of the cloud (e.g., attempting to divine, through data flows, where certain

---

<sup>26</sup> This is not meant to be a primer on how to decide whether the cloud is right for a particular organization or organizational function. For guidance on those issues, see <http://www.microsoft.com/en-us/cloud/default.aspx> and "Security Guidance for Critical Areas of Focus in Cloud Computing, v2.1" (<https://cloudsecurityalliance.org/wp-content/uploads/2011/07/csaguide.v2.1.pdf>).

<sup>27</sup> See CSIS, "A Human Capital Crisis in Cybersecurity," <http://csis.org/publication/prepublication-a-human-capital-crisis-in-cybersecurity>.

<sup>28</sup> See <http://www.informationweek.com/news/government/cloud-saas/231902850> (government cloud advocates noting the benefits of cloud security, including the increased ease of patching).

<sup>29</sup> More entities are beginning to embrace more robust authentication mechanisms. For example, Germany has issued an eID card and the Canadian Government will soon accept electronic IDs issued by banks. See <http://silicontrust.wordpress.com/2010/11/01/germany-introduces-national-eid-card-the-beginning-of-a-new-application-era/> (Germany); <http://www.finextra.com/news/Fullstory.aspx?newsitemid=23132> (Canada).

customer data may reside), but they may also use the power of the cloud to attack others.<sup>30</sup> Finally, the proliferation of devices and increased connectivity will create many ways for attackers to access systems and data without authority.

For example, someone could compromise a device to attack the cloud, or compromise the cloud to attack a device. Simply put, this new environment will require the development of new threat models which must, in turn, inform comprehensive programs to prevent, detect, contain, and recover from security events.

### III. Privacy

#### A. Background

While security was clearly the dominant pillar when the Trustworthy Computing Initiative was announced in 2002, the initial TwC communication recognized that user concerns regarding privacy were also going to be critically important to developing higher levels of trust in information technology. Indeed, privacy has been a core pillar of Trustworthy Computing from the outset, with Microsoft investing heavily to enhance and build out its privacy program. Microsoft was one of the first companies to appoint a chief privacy officer and currently has hundreds of employees responsible for helping to ensure that privacy policies, procedures, and technologies are applied across all products, services, processes and systems across the globe. These individuals are a highly diverse group, ranging from privacy attorneys who help us interpret and remain compliant with privacy rules and regulations to a mix of software engineers, marketers, scientists and business analysts who help us develop and implement our privacy program. Additionally, through our research into and development of new privacy-enhancing technologies, Microsoft has created a climate in which our employees have deep awareness of and respect for privacy issues. These investments help foster opportunities for our engineers to create technologies, services, and features that are both based on customer needs and sensitive to customer privacy concerns.

A concrete example of this work relates to the way we design and build software. Privacy standards are built into our Security Development Lifecycle and, like the SDL itself, have been made public so that all engineers can build more trustworthy and privacy capable products and services.<sup>31</sup> Our privacy principles, which we formalized and released publicly in 2008, guide the collection and use of customer and partner information and give our employees a clear framework to help ensure we manage data responsibly.<sup>32</sup> Additionally, meeting

---

<sup>30</sup> See <http://techflash.com/seattle/2011/05/sony-attack-launched-from-amazon-cloud.html> (Report: Sony PlayStation attack launched from Amazon cloud).

<sup>31</sup> See <http://www.microsoft.com/privacy/processes.aspx>.

<sup>32</sup> See <http://www.microsoft.com/privacy/principles.aspx>.

our privacy commitments is a requirement for all products and services; this is supported by clear privacy requirements that translate our global policy into actionable procedures and processes for our engineers, sales and marketing employees. We have also strengthened our support for privacy through the application of the principles of Privacy-by-Design (PbD). For Microsoft, PbD means the development and implementation of principles, policies and procedures that drive privacy-specific design objectives for our software products and online services at the outset of development and continue to address privacy and data security considerations through the product lifecycle.

## B. The Cloud and Big Data

Although our commitment to privacy remains strong, new computing models are challenging traditional efforts to protect privacy. Over the last forty years, informational privacy has been protected, in large part, by the adoption of and adherence to Fair Information Principles (FIPs). In many parts of the world, FIPs or derivatives have been codified into laws and regulations while in other parts they serve as part of self-regulatory schemas. While FIPs may vary slightly between regions of the world, they have evolved to focus heavily on certain common themes. The most important of these have included collection limitation, notice, and choice (often expressed through “consent”), redress, and security.<sup>33</sup> Certain of these principles have, over time, received greater weight. For example, there has historically been greater focus on both “notice and consent” and “collection”; the former principles supported user “control” over how information about them is collected and used, and the latter principle provided prophylactic protection since data never collected could not possibly be misused.

While perhaps conceptually satisfying, this model – with its heavy emphasis on notice and choice (consent) at time of collection – is under considerable strain and the burden it has put on individuals is untenable. Indeed, some have argued that it has virtually collapsed and is no longer serving society in the way that it was intended.<sup>34</sup> The cloud enabled world is already marked by a proliferation of devices, an abundance of data (e.g., user created content, network generated data such as geolocation and transactional data, analytical data derived from user actions), increased storage, and better search and decision enhancing algorithms.

This data rich world – which can reveal much about a person’s past and perhaps his or her expected future – presents both individuals and society with both benefits and risks. For example, analyzing data may reveal that a person has an existing medical condition that can be treated, or is at risk of developing such a

---

<sup>33</sup> See generally, Federal Trade Commission Fair Information Practice Principles (<http://www.ftc.gov/reports/privacy3/fairinfo.shtm>); OECD Privacy Principles; [http://www.oecd.org/document/18/0,3746,en\\_2649\\_34255\\_1815186\\_1\\_1\\_1\\_1,00.html](http://www.oecd.org/document/18/0,3746,en_2649_34255_1815186_1_1_1_1,00.html); and Fair Information Practices, U.S. Department of Health, Education and Welfare 1973 (reprinted at <https://www.privacyrights.org/ar/fairinfo.htm#1>).

<sup>34</sup> See “The Failure of Fair Information Practice Principles,” Fred Cate, [http://papers.ssrn.com/sol3/papers.cfm?abstract\\_id=1156972](http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1156972).

condition in the future. That data, when combined with other data, may enable startling and much-needed medical breakthroughs that benefit both the individual and society as a whole. It is important to appreciate how powerful data analytics can be. By way of example, a hospital analyzed its data to determine why certain patients were re-admitted within 30 days of discharge. Instead of asking “pre-conceived questions” (e.g., did these patients suffer from the same disease, did they undergo the same treatment), computers were asked to simply crunch the data quickly and identify interesting patterns. In this case, they discovered that patients assigned to a particular room were more frequently re-admitted with an infection.<sup>35</sup> On the other hand, it is problematic if that medical history – or information that suggests the risk of a future condition – is used to deny employment or insurance. This being the case, it can be said that a data-centric world will (1) unlock important individual and societal opportunities and (2) expose individuals and society to data use violations that are of great concern.

The real challenge, of course, is that individuals and society must reaffirm, redefine and/or fundamentally alter their expectations of privacy in this new world, particularly since the interests of data subjects and data users are not always well aligned. Indeed, as this new data-centric world has unfolded, there has been increased concern over privacy, often expressed as either the fear that privacy is dead<sup>36</sup> or the concern that there will be specific and articulable privacy harms which society must take steps to avoid. In this new world, it seems increasingly clear the privacy challenges posed by a data-centric society cannot be addressed adequately by traditional privacy principles which focus heavily on the collection of data and the notices provided at the time collection occurs.

There are three problems with this traditional approach. First, at scale, the burdens of understanding and managing choices regarding data collection and use create an enormous strain for individuals. Leaving aside that even the entity collecting the data may have a reasonable large set of understandable uses (data may be used not just to fulfill a customer’s request, but to prevent fraud or ensure the security of the network), increasingly complex business relationships invalidate the notion that a single entity will be collecting and using the data provided and that an individual and data collector will be in a strictly bilateral relationship. In fact, because an entity collecting data may partner with others or share it with numerous other organizations that provide related services (e.g., marketing companies, data analytic companies, advertising networks), individuals are confronted with an ever-increasing number of linked privacy statements to parse, providing individuals with an increasingly less practical and effective way to ensure meaningful user choice. It also turns out that, in practice, the choice offered may not provide individuals with the level of control desired; an individual must either agree to the uses by clicking on a dense privacy notice at time of collection or opt-out of participation in the desired activity.

---

<sup>35</sup> Information provided by Peter Neupert, then Corporate Vice President of Microsoft’s Health Services Group.

<sup>36</sup> Perhaps the most famous characterization of this sentiment is attributable to former Sun Microsystems’ CEO Scott McNealy who said, “You have zero privacy anyway. Get over it.” [http://www.pcworld.com/article/16331/private\\_lives\\_not\\_ours.html](http://www.pcworld.com/article/16331/private_lives_not_ours.html).

Second, the existing model assumes an interactive relationship between the individual and the entity collecting and using the data, a relationship that may not actually exist. For example, as recently reported, insurance agencies may look at Facebook photos to see if individuals claiming disability benefits are engaging in activities that suggest the insurance claim is fraudulent.<sup>37</sup> Similarly, facial recognition technology may involve matching a photo of a person in a public space with other photos of people in other public places. The actual identity of the person being “identified” may, in fact, be unknown at the time of the data collection or usage, thus making discussions about user consent impractical. In such a case, an organization collecting and using the data may be making decisions about an individual with whom they have no relationship or even knowledge of identity.

Third, the true value of data may not be understood at the time of collection and future uses that have significant individual and societal benefit may be lost. This is not to suggest, of course, that unneeded data should be collected (which, in any event, leads to higher storage and security costs) or that data collected for one purpose should be used cavalierly whenever some new, beneficial purpose is later discovered. Rather, the point is that this new data-centric world will provide society with a wide range of new possibilities, some of which may be unforeseen at the time data is collected (e.g., those collecting blood samples forty years ago did not mention DNA testing as a potential use, but it has since served to exculpate the innocent). At the same time, it is certainly true that data can be abused; redlining (where loans are denied to neighborhoods comprised heavily of minorities) presents one such example. But whether data usage is “good” or “bad” is just that – a judgment about usage. While a limitation on data collection does serve an important prophylactic purpose and remains relevant, it is not without cost; the real debate should be over use and how to strike a right balance between the societal benefit and personal privacy. That has been, is, and will likely remain an issue requiring deep thought.

If changes in the way we apply FIPs to a data rich world are necessary – and clearly they are – the new focus must take into account a new world that is ripe with new business models, new data usage models, new forms of technology, and individual privacy sensibilities that may prove to be either remarkably resilient over time or quite fluid. Already today, some would argue that Internet interactions are evolving into barter-like transactions where individuals exchange personally identifiable information for services, while others might argue that this bartering is not very transparent in terms of the “value” being traded (i.e., it is not clear what a filled-in gender check box is worth, nor what one gets in return for checking the box). Still others might argue that personal information should be transparently “priced” – and then bought and sold – like any other asset.<sup>38</sup>

---

<sup>37</sup> “Are insurance companies spying on your Facebook page?”, <http://abclocal.go.com/kabc/story?section=news/consumer&id=8422388> (Monday, November 07, 2011).

<sup>38</sup> See, for example, Start-Ups Seek to Help Users Put a Price on Their Personal Data, [http://www.nytimes.com/2012/02/13/technology/start-ups-aim-to-help-users-put-a-price-on-their-personal-data.html?\\_r=1](http://www.nytimes.com/2012/02/13/technology/start-ups-aim-to-help-users-put-a-price-on-their-personal-data.html?_r=1).

All of these arguments have merit and can be argued logically, and it remains unclear how changing privacy sensibilities about data collection and use will influence these various approaches. The two core questions to be answered during this period of change are (1) should Fair Information Principles be evolved to reflect this changing environment and, if so, (2) can such principles serve the twin goals of protecting privacy and enabling data usage for both individual and societal benefit?

Given today's data-centric world, it seems clear that the use of data, rather than its collection and associated notice and consent schema, serves as a better focal point for defining the obligations related to personal information. While it might be jarring at first glance to de-emphasize the principle relating to collection, it is important to note that collection and notice principles remain relevant in the new model; the difference is that the primary focus moves away from collection and notice and towards use. Such a use model<sup>39</sup> requires all organizations to be transparent, offer and honor appropriate choice (control), and ensure that risks to individuals related to data use are assessed and managed. This captures all uses (not just interactive uses that stem from bilateral relationships) and provides a governance approach that is more effective for individuals, more manageable for business, and permits better oversight by regulators.

In applying a model such as this, it is important to recognize that while there are new forms of data and new forms of data usage, there are, in the end, only three ways to think about data use: There are uses that are (1) broadly acceptable, and sometimes legally authorized or required; (2) prohibited; and (3) subject to individual sentiment (that is, the usage may be acceptable to some but reasonably objected to by others such that a blanket rule is inappropriate). It must also be recognized that while society or an individual may designate a use as acceptable or unacceptable, this view may change over time and it may be difficult to enforce a new usage rule on data that already exists (e.g., if a person makes information publicly available but later has remorse after posting, it may be impossible to exert meaningful control over the proverbial genie that has left the bottle). Still, by thinking of usage in this way, society can now decide how to create frameworks that permit the right uses to be placed in the right categories and, thereafter, how to craft appropriate "obligations" associated with particular uses.

For example, in the broadly acceptable category, certain uses may be deemed acceptable even absent explicit user consent. This might include using consumer data to provide the requested product or service (order fulfillment); selling or marketing related products and services (unless the consumer has opted-out of such data use); improving the service; engaging in research (so long as the data is de-identified, in appropriate cases, through business and/or technical means); preventing fraud against the service; and reporting illegal activities to those government authorities responsible for investigating them, if lawfully permitted or required. If agreement can be reached on what constitutes broadly acceptable use, individuals will come to expect these

---

<sup>39</sup> The Business Forum for Consumer Privacy – Use and Obligations – a Path Forward.

See [http://www.huntonfiles.com/files/webupload/CIPL\\_Use\\_and\\_Obligations\\_White\\_Paper.pdf](http://www.huntonfiles.com/files/webupload/CIPL_Use_and_Obligations_White_Paper.pdf).

uses and will be able to minimize time reading the notices related to these practices. With less focus on broadly acceptable uses, choices relating to uses where reasonable minds can differ can stand out more prominently and energy can be spent managing the risks associated with these activities, including by opting-in or opting-out of a given usage depending on the sensitivity of the data. Then organizations using the data can be held accountable for honoring those preferences. Note that notice should still be provided regarding use practices; the real difference is that the burden for enforcing the rules moves from the individual (who too often has to fend for himself or herself after reading linked privacy statements) to the organization using the data. This permits more consistent, workable, and enforceable societal rules to be crafted – rules that then can be better enforced by technology (e.g., by applying policies to metadata), individuals, and regulators.

Under this model, it is important that organizations using data be truly accountable for meeting their obligations. Thus, FIPs in a data-centric world should include an “accountability principle”: an entity receiving data (directly or indirectly) is responsible for ensuring that such data is collected lawfully, and used and protected in ways consistent with individual and societal expectations. To be “accountable” means that the organization has taken steps to develop and implement privacy risk assessments, policies, processes and procedures that help enforce data usage rules that honor societal norms, respect user control, and ensure data is reasonably secure.<sup>40</sup> Significantly, upon request of a regulator, an organization must be able to demonstrate how they have fulfilled their responsibility under this principle.

As is true today, those collecting data from individuals for a business purpose must provide notice indicating (1) the purposes for which that data will be used; (2) whom the data will be shared with, if anyone; (3) how such use or sharing benefits the individual; (4) whether the user will remain non-identifiable through either business process or technological means; (5) the general type and nature of the information collected, (6) what control, if any, the individual can exercise to affect the use and sharing of this data, and (7) how the data is protected. All collection and data use must, of course, meet legal requirements and be reasonably consistent with the notice provided.

Where reasonably practical, user control should be granular, contextual, and automated, thus permitting more meaningful decisions regarding the transfer of data in return for value. In this regard, some existing rules provide good guidance. For example, as part of that control, explicit consent should be obtained for the collection, transfer, and additional use of data or where the data being collected or shared is sensitive.<sup>41</sup> This

---

<sup>40</sup> What the Accountability principle would require of organizations has been generally accepted and agreed to. See

<http://www.informationpolicycentre.com/resources/> and [http://www.huntonfiles.com/files/webupload/CIPL\\_Galway\\_Accountability\\_Paper.pdf](http://www.huntonfiles.com/files/webupload/CIPL_Galway_Accountability_Paper.pdf).

<sup>41</sup> Sensitive Personally Identifiable Information (a.k.a. Sensitive Consumer Information) includes PII and pseudonymous data that could (i) be used to discriminate (e.g., race, ethnic origin, religious or philosophical beliefs, political opinions, trade union memberships, sexual lifestyle, physical or mental health), (ii) facilitate identity theft (e.g., mother’s maiden name), or (iii) permit access to a user’s account (e.g., passwords or PINs). It also includes other data that is not technically PII, but has historically made users nervous (such as a user’s precise location) and data that has a reasonable expectation of embarrassing the user.

control should also address rules about anonymous or de-identified data. For example, in situations where the value of data use is not undermined, an individual should have the option to remain non-identified, either through established business processes<sup>42</sup> or technological means where possible. Finally, traditional principles relating to access and redress, security, and, today, breach notifications, all make eminent sense in our new data-centric world and should be carried forward with rigor.

As noted, this approach represents a shift in focus – it builds on 40 years of history rather than discarding it – but is designed to help achieve five ends: (1) protect privacy in meaningful ways; (2) optimize the use of data for the benefit of both individuals and society; (3) ensure that those who use data are accountable for that use; (4) provide a regime that permits more effective oversight by regulators, and (5) work effectively in a modern connected society. In a data rich world, achieving these objectives requires a focus on data use, meaningful user control, and transparency.

### C. Government Access to Data

As individuals and organizations embrace services, questions arise about government access to data, particularly when that access comes without notice. To some extent, this new data-centric and service oriented world does shift the balance of power between the individual and the government. This is because, in the past, many personal documents (e.g., personal correspondence, photographs and other personal papers) were stored in the home or in the office, even if other records (e.g., telephone records, banking records) may have been available from a third party. This being true, a government seeking such data had to engage an individual directly, even if that engagement involved a forced seizure (e.g., the execution of a search warrant). Today, an individual using IT services may store a wealth of personal information with a third party, including correspondence (e-mail), text documents, photographs, and copies of financial and health records. Additionally, other forms of interesting information (what has been purchased online, precise location data) is also available. To the extent that governments can obtain such information from service providers without notice to the individual, it is a cause of concern, especially since it is increasingly difficult to participate fully in society without leaving a significant digital trail.<sup>43</sup>

---

<sup>42</sup> This reference to “business processes” is important, as many will argue that any de-identification of data can be overcome through re-identification of data. While our technical ability to prevent re-identification of data may be limited, technological means are not the only option available to enforce societal norms. For example, laws prohibited the interception of cellular phone calls even when, technically, achieving that end was not difficult.

<sup>43</sup> In the United States, well established court precedent holds that data provided to a third party could be obtained by the government without violating the Fourth Amendment’s search warrant requirement. See *Smith v. Maryland*, 442 U.S. 207 (1986). Recently, however, one Circuit Court held that to the extent the Electronic Communications Privacy Act permitted the Government to access e-mail with less than a warrant, it was unconstitutional. See *United States v. Warshak*, <http://www.ca6.uscourts.gov/opinions.pdf/10a0377p-06.pdf>.

There are, in fact, two distinct issues. The first relates to domestic investigations. Assuming an investigation is wholly within one country, under what rules should a government get access to an individual's data? The second issue relates to international investigations. In international cases, to what extent should a government be able to compel the production of documents that are stored in another country without seeking international assistance?

## 1. Domestic Investigations

The issue of government access to locally stored records has long been the subject of privacy discussions. As noted in one U.S. Government report:

The existence of records about an individual that are not in his possession poses serious privacy protection problems, especially when government seeks access to those records. Record keepers can, often do, and sometimes must, disclose records about an individual to government without seeking the individual's approval, whether the disclosure is at the request of government or through the initiative of the record keeper; and, frequently no record of the disclosure is ever made....The individual may never know that agents of the government have inspected his records. Except in a limited number of situations, neither the record keeper nor the government is obliged to notify him that his records were opened to government scrutiny.

This report was not about big data and cloud services; in fact, it was written before the age of personal computing.<sup>44</sup> But as individuals increasingly store data in the cloud (and therefore in the possession of a third party), to what extent should that fact erode a person's expectation of privacy? One U.S. Supreme Court Justice has recently commented on this issue:

More fundamentally, it may be necessary to reconsider the premise that an individual has no reasonable expectation of privacy in information voluntarily disclosed to third parties. *E.g., Smith*, 442 U. S., at 742; *United States v. Miller*, 425 U. S. 435, 443 (1976). This approach is ill suited to the digital age, in which people reveal a great deal of information about themselves to third parties in the course of carrying out mundane tasks. People disclose the phone numbers that they dial or text to their cellular providers; the URLs that they visit and the e-mail addresses

---

<sup>44</sup> Chapter 9 of *Personal Privacy in an Information Society: The Report of the Privacy Protection Study Commission* transmitted to President Jimmy Carter on July 12, 1977, reprinted at <http://aspe.hhs.gov/dataacncl/1977privacy/c9.htm>.

with which they correspond to their Internet service providers; and the books, groceries, and medications they purchase to online retailers. Perhaps, as JUSTICE ALITO notes, some people may find the “tradeoff” of privacy for convenience “worthwhile,” or come to accept this “diminution of privacy” as “inevitable,” *post*, at 10, and perhaps not...I would not assume that all information voluntarily disclosed to some member of the public for a limited purpose is, for that reason alone, disentitled to Fourth Amendment protection.<sup>45</sup>

Even before the Jones decision, Microsoft itself had asked that the standard for seeking data from third parties be reconsidered.<sup>46</sup>

The advent of cloud services and big data clearly exacerbate such privacy concerns. While data has long existed that reflects, at a high level, a person’s activities (e.g., credit card charges leave a trail of what was purchased, including when and where), the quantity and quality of revealing data is increasing dramatically. Today, instead of cobbling together a paper record or relying upon physical surveillance to track a person’s location, pictures with metadata such as date and time may be taken by public cameras and later analyzed with facial recognition technology. Very precise GPS signals can be captured from a range of mobile devices including phones, cars, and computers. In this new environment, one can readily recognize how helpful this data may be to those investigating unlawful activity<sup>47</sup> or, for that matter, governments attempting to chill lawful activity related to freedom of association. Clearly, the time has come for a robust debate about the protection of information stored in the cloud.

## 2. International Investigations

While countries can of course decide what rules apply domestically, how governments will access data stored in foreign countries raises far more complex questions. Simply put, as more people move data to the cloud for anywhere/anytime access, governments all over the world will want access to that data, raising serious questions that lie at the intersection of communications, geography and, ultimately, sovereignty. In some parts of the world, the expressed concern is “the Patriot Act,” a U.S. law passed after the terrorist attacks of 9/11 that, it is claimed, gives the U.S. government the ability to compel U.S. companies to produce data in their possession, even if that data is stored in other countries.<sup>48</sup> In fact, claims that the Patriot Act is relevant are misguided; the

---

<sup>45</sup> See *United States v. Jones*, <http://www.supremecourt.gov/opinions/11pdf/10-1259> (Justice Sotomayer, concurring).

<sup>46</sup> See Brad Smith, “Cloud Computing for Business and Society,” at [http://www.huffingtonpost.com/brad-smith/cloud-computing-for-busin\\_b\\_429466.html](http://www.huffingtonpost.com/brad-smith/cloud-computing-for-busin_b_429466.html).

<sup>47</sup> In the United States, the Supreme Court recently held that government agents must have a search warrant before placing a GPS device on a vehicle. See *U.S. v. Jones*, <http://www.supremecourt.gov/opinions/11pdf/10-1259>.

<sup>48</sup> See <http://www.infoworld.com/d/security/european-distrust-us-data-security-creates-market-local-cloud-service-180706> (“European distrust of U.S. data security creates market for local cloud service; Europeans worried about the U.S. Patriot Act prefer to keep their data in the EU”).

U.S. legal precedent regarding data access was established almost two decades before the terrorist attacks of 9/11 and, therefore, long before the Patriot Act was passed. As many are unfamiliar with this history, it is worth sharing.

On March 4, 1983, the Bank of Nova Scotia, a Canadian bank headquartered in Toronto, had branches, offices and agencies in forty-six countries. On that day, the Bank's Miami, Florida, office was served with a grand jury subpoena duces tecum issued by the United States District Court for the Southern District of Florida. The subpoena called for, among other things, the production of financial documents pertaining to two individuals and three companies from the Bank's branch in the Cayman Islands. The bank claimed that it could not produce the requested documents because doing so would violate Cayman Islands secrecy law. The district court issued an order compelling the bank to produce the documents, a decision ultimately upheld by the appellate court:

In a world where commercial transactions are international in scope, conflicts are inevitable. Courts and legislatures should take every reasonable precaution to avoid placing individuals in the situation [the Bank] finds [it]self. Yet, this court simply cannot acquiesce in the proposition that United States criminal investigations must be thwarted whenever there is conflict with the interest of other states.

Simply put, the court held that the United States government can compel a U.S. located company to produce data in its possession, no matter where that data is stored.<sup>49</sup> Interestingly, no other government has taken a contrary view or ever suggested that it would refuse to compel the production of documents from a local company simply because the data was stored in a foreign country.

It is an interesting exercise to replace the word "bank" with "cloud provider" and then think about the implications for cloud users, cloud providers, and governments. For users, the question is, "what governments can get my data and will I know that it has been released?"<sup>50</sup> Users may understandably be concerned that foreign governments can access their data. Additionally, which governments can access that data may change if, for example, a service provider established a presence in yet another country or, for reliability's sake, backed up a user's data to a new geographic location. Cloud users also cannot be assured that they will receive notice if a governmental entity seeks their data because, while cloud providers may have policies requiring such notice,

---

<sup>49</sup> See *Bank of Nova Scotia v. United States*, 740 F.2d 817, 828 (11<sup>th</sup> Cir. 1984), cert denied, 469 U.S. 1106 (Jan. 7, 1985). As *Bank of Nova Scotia* makes clear, such access was well-established law long before 9/11 and, interestingly,

<sup>50</sup> As noted, all governments have the right to compel the production of data from entities located within their jurisdiction. That said, because many major cloud providers have a presence in the U.S., the question of U.S. government access is often a focal point of discussion.

governments can serve non-disclosure orders that prohibit such disclosures. That being the case, how should cloud users think about this issue?

Government access to data is simply one risk management factor to be considered when thinking about the use of cloud services. For example, an organization may want to avoid putting highly sensitive trade secret information in a cloud service if doing so will permit a government that lacks respect for intellectual property rights to get access to that data. At the other end of the spectrum, using a cloud service to collaborate on data that will be made publicly available may raise no concerns at all. It is also important to note that many organizations have migrated to the cloud in part because the cloud provider may offer more effective security controls than are implemented on premises. In such cases, an organization could reasonably conclude that the business risks of a hacker stealing sensitive information from an on premises system is greater than the risk of a government gaining access to that data through legal process.

It is also important to quantify the risk appropriately. When a multi-national company has a presence in a given country, it is already subject to that country's jurisdiction. As such, the government can compel the production of the company's documents, regardless of where those documents are stored. In such cases, the risk of using cloud services is not that the data can be accessed by the government (the government can access it anyway) but that such access will occur without notice (that is, the government can get access to the data through the cloud provider and prohibit notice to the company under investigation). The risk of access without notice must also be weighed appropriately because, in many law enforcement investigations, companies are well-aware of a government's investigative activities: people are interviewed; closely held documents are reviewed; and companies even issue public statements that they are cooperating with investigators. To the extent that a customer routinely cooperates with government investigations, it may matter little whether the documents are held by the company on premises or in the cloud. In summary, it is important for cloud users to ask the right questions and balance the risk of government access with the productivity and other benefits enabled by the cloud.

Bank of Nova Scotia also has implications for cloud providers. The key question is, "how does a cloud provider comply with all of the laws in all of the countries that can compel the production of data, especially if complying with the law in one country violates the law of another?" Legitimate cloud providers have no interest in violating national laws, but sovereigns may nonetheless put them in this untenable position because national laws differ and, as Bank of Nova Scotia makes clear, governments are unlikely to voluntarily limit their investigative authority.

It is also worth noting that Bank of Nova Scotia involved a conflict between a criminal investigation in one country and a civil law in another. In the future, there may well be situations where criminal laws come into conflict and a company's employees are told that whatever decision is made, they will be subject to criminal

prosecution and, possibly, imprisonment. While the court in the Bank of Nova Scotia case may be correct that it is unworkable to allow organizations to avoid local laws by storing data in foreign countries, it is also unworkable – and it is inherently problematic – to force law abiding people to break criminal laws, no matter what their behavior.

As for governments, the questions might be, “under what authority do I get what data”; “how do I get access to data when neither the data nor cloud provider is within my jurisdiction, even though all the criminals – and the crime – may be”; and “how do I invoke my traditional powers – which may allow access to both stored data and communications in transit – in this new cloud environment”? Making the answers to these questions even more difficult is that a government engaging in an online investigation may not know the true identity or current location of an investigative target, nor be able to identify with certainty the jurisdiction in which that person’s data is stored. As such, investigative rules dependent on geography, nationality, or other readily determined factors have proved challenging.<sup>51</sup>

The economic and societal benefits of cloud computing and big data – benefits governments support – arise in part from data consolidation in large data centers that are globally dispersed; locating a data center in every country and having each data center serve only the local population undermines this model. At the same time, however, governments will not waive their right to compel data whenever they can, in part because they cannot fulfill their responsibility to protect public safety and national security without access to such data.

While denying a government access to critical data, even when they have a court order, cannot be right; neither can it be right to make all the data of all cloud users available to every government without a legal regime that protects the civil liberties and other interests of affected parties. Thus, it is important for governments to focus on creating more workable rules for the future and ask whether there is any framework that may alleviate this conflict, at least in some set of cases. Indeed, the fact that some cases may be uniquely difficult (e.g., cases involving important differences in law and culture, or cases where governments are accusing each other of misconduct) does not mean that all cases must be difficult; rules that would work in many cases would serve to reduce the number of international conflicts considerably.

Under a new framework, one could start with the view that the country where data sits has jurisdiction over that data. Countries could agree that, at least for some offenses, a requesting country can use its judicial processes to demand data from a company doing business within its borders, even if that data is stored

---

<sup>51</sup> Even if location is known, international mechanisms for cross-border assistance do not move at the speed of bits and, more importantly, are voluntary (mutual legal assistance treaties may permit and expedite assistance, but do not require it). Additionally, in some cases, national laws prohibit assistance. For example, the United States wiretap statute permits wiretapping only if a U.S. agent alleges a violation of certain U.S. laws (see 18 U.S.C. 2516 for a list of designated offenses). To the extent someone in a foreign country is using a U.S. electronic communications service to engage in criminal activity in that foreign country, the U.S. has no legal ability to provide wiretapping assistance. Finally, some governments have questioned why they must rely upon international assistance when seeking electronic evidence regarding local people who have allegedly violated local laws.

elsewhere. The data requested would have to relate to individuals in the requesting country's jurisdiction or to crimes that are/will occur in that jurisdiction. As part of this framework, the compelled company would notify the data sovereign (the country from which the data is being pulled) of the nature of the request (what is being sought and by whom). The purpose of this notification would be to ensure that if the data sovereign has concerns about the request, those concerns can be discussed with the requesting sovereign. It is true that some countries may be concerned about disclosing the existence of their investigation, but if the framework is scoped appropriately (e.g., to offenses of mutual concern), such a disclosure should not be problematic. Finally, it would be agreed that entities responding to compulsion orders under this framework would be acting in accord with international law and not subject to prosecution.

## IV. Reliability

### A. Background

In 1999, Microsoft Research formed the Programmer Productivity Research Center (PPRC) to look at various techniques to improve code quality. The team developed new static analysis and testing techniques, as well as tools to help improve and further automate software quality assurance. For example, PPRC developed static analysis tools for C/C++, dependency analyzers, code coverage tools, test prioritization tools, performance analyzers, and advanced debuggers. These tools have played key roles in improving software quality for Microsoft products, particularly for Windows and Office. This effort increased in intensity when TwC was announced, and Microsoft engaged in additional efforts to embrace practical, measurable reliability definitions and use data provided from increasingly instrumented products to determine if its reliability objectives were being met. Of course, defining reliability – and achieving it – is no easy matter. From a definitional standpoint, it has been noted that “The literature of both the academic and commercial worlds is replete with discussions of ‘reliability.’ We define it, we argue about it, we compare implementation against idealized standards of it.”<sup>52</sup> Nor are dictionaries much help. Merriam-Webster defines “reliability” as “the quality or state of being reliable” or, more specifically, “suitable or fit to be relied on: dependable.”<sup>53</sup> It turns out that this definition is unsatisfying to customers who may have different expectations of what it means for a product or system to be reliable. Some focus on availability, some include performance measures, others add manageability and recoverability, and some focus on predictability.

Although it is not possible to harmonize the many diverse views of reliability in this paper, we do need to be clear about its use in the context of TwC. In a world of personal computers, it has been said that “for engineering purposes, reliability is defined as: the probability that a device will perform its intended function

<sup>52</sup> Thompson, Herbert and Whittaker, James, “Understanding Reliability: Measuring IT Pain Points” (May 2005).

<sup>53</sup> See <http://www.merriam-webster.com/dictionary/reliability>.

during a specified period of time under stated conditions.”<sup>54</sup> Even this simple definition posed certain unique challenges in the world of IT; unlike cars and traditional telephones, those using IT systems found they were widely adaptable and could use them in ways that varied widely from any “stated conditions.” Perhaps more importantly, this traditional definition focuses on the device, and, in a world of increasingly connected devices and services (and the data they contain), this definition fails to take into account whether the service actually meets objectively reasonable expectations of performance. In a connected world where we depend on devices and services for a range of critical and non-critical activities, reasonable user expectations related to reliability will need to be met before people declare a system trustworthy.

This is not as simple as it sounds, and we cannot simply rely upon historical efforts which emphasized preventing faults (an inherent weakness in design or implementation) that led to failures, often expressed as “mean time to failure” (MTTF). While it remains important that each device be reliable (since a service is rendered by a combination of devices), services are increasingly taking on both known and unknown technical and organizational dependencies. This is happening at a time when there are an increasing number of natural disasters challenging the survivability of systems and the availability of components, and the recovery of impaired facilities can be protracted.<sup>55</sup>

To appreciate the complexity of our current world, one need only contrast an old and new world failure. When house lights went out in the old world, homeowners engaged in a simple response: they would look out the window to see if the rest of the neighborhood was dark. If lights were on elsewhere, the homeowner had an issue; if lights were out everywhere, the power company had an issue. Even if the issue was with the power company’s complex electric generation and delivery system (power plants, transmission lines, substations, and the connection to the home), it was mostly within the confines of the power company to identify and repair the issue. Put another way, failures might have technical dependencies but not organizational dependencies; a single entity controlled most of the infrastructure.

Today, systems are far more complex, with both technical and organizational dependencies. If a user cannot access a webpage from a laptop in one’s home, it might be (1) the laptop (hardware, software, application); (2) the wireless router; (3) the broadband modem; (4) the access provider (transmission line or back-end system); (5) a remote system (hardware, software, application) or (6) an actual power failure somewhere along that chain of connectivity. While diagnostic tools do exist, identifying and resolving problems is often challenging.

---

<sup>54</sup> [http://en.wikipedia.org/wiki/Systems\\_reliability](http://en.wikipedia.org/wiki/Systems_reliability), citing, in part, (Citing Institute of Electrical and Electronics Engineers (1990) IEEE Standard Computer Dictionary: A Compilation of IEEE Standard Computer Glossaries. New York, NY ISBN 1559370793).

<sup>55</sup> “The number of disasters has grown from fewer than 100 in 1975 to more than 400 in 2005, and increasing at a steady rate each year.” EM-DAT: The OFDA/CRED International Disaster Database – [www.emdat.be](http://www.emdat.be), Université Catholique de Louvain, Brussels (Belgium). See also <http://bits.blogs.nytimes.com/2011/11/04/thailand-floods-will-affect-computer-makers-and-web-sites/>.

## B. The Cloud

The complexity of these global systems is certainly not lost on those thinking about cloud adoption,<sup>56</sup> concerns exacerbated by recent, highly publicized outages suffered by virtually all cloud service providers.<sup>57</sup> It is not merely a matter of inconvenience; the reliability challenges of IT systems may affect business productivity<sup>58</sup> or even public health and safety. In times of crisis, governments may use social media to keep citizens informed, and first responders may be empowered to react effectively not just because they have radios, but GPS devices, mapping capabilities, street views, video conferencing, and other cloud based services. Such benefits only materialize, however, if information systems meet reasonable expectations of overall service reliability. Recognizing this fact, governments are increasingly looking at the cloud – or at least certain cloud components – as part of critical infrastructure.

To the extent the Internet was not designed to be secure, the reliability picture is more complex. On one hand, the Internet was designed to withstand military attacks; that is, it was designed to be dependable even in the most stressful of times. That said, as with security, it was expected that use of the Internet would be limited to particular purposes and that those using the capabilities of the Internet would be trustworthy. It was not designed for the number of users and variety of uses in play today, nor was it anticipated that it would be used to deliver malicious packets. Considering its intended use and intended user base, a “best effort” delivery system met the necessary requirements of the time. But as technology is increasingly embedded in the fabric of our lives – in situations ranging from social interaction to commercial transactions to emergency response – expectations regarding reliability have risen and we must achieve a level of reliability which is not a given today.

Indeed, there has been increasing recognition that the interdependencies created, particularly among critical infrastructures, is a cause for concern:

Our national defense, economic prosperity, and quality of life have long depended on the essential services that underpin our society. These critical infrastructures—energy, banking and finance, transportation, vital human services,

---

<sup>56</sup> When asked to summarize why they are not planning on using cloud services, IT professionals noted that security, privacy and reliability are primary concerns. ISACA/ITGI “Global Status Report on Governance of Enterprise IT”, January 2011, (pp. 38)

<sup>57</sup> Outages have affected, among others, Microsoft ([http://windowsteamblog.com/windows\\_live/b/windowlive/archive/2011/09/20/follow-up-on-the-sept-8-service-outage.aspx](http://windowsteamblog.com/windows_live/b/windowlive/archive/2011/09/20/follow-up-on-the-sept-8-service-outage.aspx)), Amazon (<http://aws.amazon.com/message/65648/>), Google (<http://informationweek.com/news/cloud-computing/software/231600978>), and VMware (<http://support.cloudfoundry.com/entries/20067876-analysis-of-april-25-and-26-2011-downtime>).

<sup>58</sup> See IT Downtime Costs \$26.5 Billion In Lost Revenue, [http://www.informationweek.com/news/storage/disaster\\_recovery/229625441](http://www.informationweek.com/news/storage/disaster_recovery/229625441) (noting that IT failures can also rattle the confidence in new technologies such as cloud computing).

and telecommunications—must be viewed in a new context in the Information Age. The rapid proliferation and integration of telecommunications and computer systems have connected infrastructures to one another in a complex network of interdependence. This interlinkage has created a new dimension of vulnerability, which, when combined with an emerging constellation of threats, poses unprecedented national risk.<sup>59</sup>

Although these interdependencies are recognized, it remains very difficult to manage risks to reliability because entities do not always have clear visibility into technical or operational dependencies. An illustration may be helpful. Banks may rely upon “redundant” communications channels, from different service providers, to ensure the availability of the electronic funds transfer network. At the same time, different service providers may share “hotel space”; their communication lines may run through a shared facility to save costs. This means, of course, that if the facility is damaged, the bank loses both channels of communications at the same time (it is a single point of failure). There is no easy way to know that the intended redundancy was not created, since the bank may not share its network map with the telecommunications companies, and the telecommunications companies may not share their maps with the customer or each other.

What, then, must change? In short, there are two fundamental changes that must occur. First, just as big data can be useful in predicting human behavior, we need to leverage big data to create “engineering intelligence” (EI): using “big data” to identify, extract and analyze large amounts of engineering related data throughout the lifecycle of a product, including in-development and in-operation, to enable better engineering related decision-making and thereby improve overall engineering quality and productivity. In that sense, EI technologies can provide historical, current and predictive views of engineering and operations. For example, simply watching data flows between networks may reveal significant dependencies which were previously not understood. This is an area ripe for research and tool development. We must also create a workable taxonomy to assess criticality from a dependency standpoint. While vulnerabilities in products have been assessed and mitigated based on ease of exploit and risk to the customer, mitigating vulnerabilities introduced by dependencies across services or across companies is a harder challenge, and one that also requires greater thought.

Second, in the shorter term, we need to rethink the way products and services are engineered to ensure resiliency. Historically, availability improvements related to systems have been achieved by improving the quality of individual components, and through redundancy and data replication. Redundancy was designed to ensure the duplication of critical elements such that the failure of one element would not interrupt the operation of the system. While undoubtedly effective at addressing a wide range of potential failure modes, simple redundancy mechanisms are proving to be insufficient in terms of assuring high levels of reliability in the

---

<sup>59</sup> Marsh Commission Report, “Critical Foundations, Protecting America’s Infrastructures”, p. ix., found at <http://www.fas.org/sgp/library/pccip.pdf>.

cloud. While software and hardware suppliers invest heavily in designing robust failover mechanisms, these failover mechanisms are not proving up to the task. Leaving aside that they must be properly implemented, properly maintained and – over time – upgraded without interruption while in active use, the current failure rate of large scale systems suggests these past efforts are not sufficient.

Data replication was also broadly adopted by IT professionals and individual users to guarantee data is readily available and consumable in the event of device failures. Having multiple identical copies of the data distributed amongst different devices and hosted in different locations reduces the likelihood of data loss and thereby improves overall system reliability, but it also introduces a level of complexity in terms of how the systems must handle updates to preserve the consistency of the data across these different devices and/or locations. An application-level malfunction or a mistake made during a maintenance activity can result in the unintentional corruption or loss of data. If a malfunction or mistake goes undetected for some period of time, the data replication mechanism responsible for creating exact duplicates of the source can corrupt those critical duplicates as well.

While redundancy and replication are important, more must be done. Software and hardware vendors must look at reliability from two perspectives. First, instead of focusing primarily on the reliability of a given component, it is important to consider how that component contributes to the overall reliability of today's more complex ecosystem. In this regard, fault modeling should be used early in the software lifecycle to achieve the most benefit. Second, through collaborative efforts, the industry can map out end-to-end scenarios for "composite reliability" and work together on architectures and standards for production implementations that – taken as a whole – provide end-to-end reliability. It may be useful to build industry organizations that can help define the high priority scenarios and drive the development of best practices and standards – a Cloud Reliability Alliance analogous to the Cloud Security Alliance, for example.

Second, in light of our dependence on the cloud and the increased complexity of the environment, the historical emphasis placed on preventing failures in software needs to be supplemented by an increased focus on software that detects, isolates, and repairs (or works around) the routine failures associated with composite computing systems. Indeed, there are several important factors unique to cloud services that compel the service designer or service operator to devote far more time and effort toward what has been termed "recovery-oriented computing."<sup>60</sup> Routine faults in the computing environment supporting a cloud service, regardless of whether they manifest themselves in the form of device failures, latent vulnerabilities in software or firmware, or human errors, are unavoidable; thus, software should expect these conditions and be designed for failure. The dependencies between the service components should be as loosely coupled as possible, and each component should react to other components' failures by gracefully degrading, providing a partial service experience rather than creating a service down condition. This implies the need for the cloud service designer to create fault

---

<sup>60</sup> See [http://roc.cs.berkeley.edu/roc\\_overview.html](http://roc.cs.berkeley.edu/roc_overview.html).

models at design time not just the traditional component fault modeling referred to above, but true end-to-end fault modeling.

The designers must also verify that the “coping mechanisms” described in the specification are reflected in the software being developed and then tested through fault injection in the actual production environment to confirm that the expected behavior materializes in the real world (test environments are proving increasingly unhelpful in a world of global large-scale services). Ideally, this deliberate injection, now referred to as “test in production,” would be done programmatically to continuously verify that subsequent releases of software, changes in network capacity or design, and/or the addition of new subsystems have not, in effect, introduced a heretofore undetected reliability threat.<sup>61</sup> These efforts must also include the ability to seamlessly roll back changes when faults are identified, with a level of reliability precluding actual users of the services from being negatively impacted.

## V. Conclusion

When Bill Gates announced Trustworthy Computing, computing and society were at a major inflection point. Our increasing dependence on IT systems brought into sharp relief the importance of focusing on the security, privacy, and reliability of software products. Today, we are at another inflection point. Computing is marked by a myriad of devices, global services, and big data. As great as our dependency on computing was in 2002, it has grown dramatically in the last decade. The Internet, long a vehicle for educational advancement and commercial growth, now stitches together the social fabric of society, even playing a major role in democratic revolutions that have marked recent history.

It has long been said that the only constant is change and, as the world’s relationship with computing continues to evolve, TwC must evolve as well. It is worth noting that, even in hindsight, the work of the company over the last decade has been critically important. Many have embraced the Security Development Lifecycle, our relentless focus on privacy has served our customers well, and our efforts in reliability have largely relegated system crashes to the historical dustbin. But in a world marked by complete dependence on information technology, determined and persistent adversaries, a proliferation of data, devices, and services, and governments concerned about protecting users, the Internet, public safety and national security, the strategies we have formulated to protect security, privacy and reliability must continue to evolve.

---

<sup>61</sup> The canonical example of deliberate, programmatic, fault injection is best represented by the so-called “Chaos Monkey” tool developed by Netflix, which has since been expanded to cover a wider range of potential failure conditions, (now referred to by Netflix as the “virtual Simian Army”). See <http://techblog.netflix.com/2011/07/netflix-simian-army.html>.

In each substantive pillar of TwC, we confront new and unique challenges, and we must rise to meet them. By adopting a more holistic security strategy that encompasses prevention, detection, containment, and recovery, the world can better address increasingly determined and persistent adversaries. By understanding what it means to live in a highly connected, device-laden and data rich world, we can craft principles that remain effective in protecting privacy but allow us to reap the benefits that only big data can bring. By leveraging engineering intelligence and pursuing recovery oriented computing, we can create products and services that are flexible in times of failure and help ensure the reliability of devices and services notwithstanding the complexity, interconnectedness, and dependencies that exist in our information systems. Finally, by being open and transparent in our business practices, we can engender the trust of those dependent on information technology. In sum, the mission defined by Bill Gates ten years ago remains as vital and important as ever.

## 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 the concepts 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.

With that caveat, I particularly want to thank the following contributors: Matt Thomlinson, Adrienne Hall, Fred Schneider, Jeannette Wing, Steven Gerri, Neeraj Suri, Malcolm Crompton, Dean Hachamovitch, Martin Abadi, Mark Russinovich, Fred Cate, Ellen Cram Kowalczyk, Dan Reed, Scott Field, Peter Haynes, Steve Lipner, Vijay Varadharajan, Viktor Mayer-Schonberger, Mike Adams, Xuedong Huang, Jeff Jones, Diane D’Arcangelo, Shawn Aebi, Reese Solberg, Ellen McDermott, Peter Loforte, Rich Wallis, Cristin Goodwin, Geff Brown, and Adam Shostack. Also, I offer a very special “thank you” to Peter Cullen, Brendon Lynch, Jules Cohen, and David Bills for their extensive and unique contributions.

Trustworthy Computing Next

© 2012 Microsoft Corp. All rights reserved.

This document is provided "as-is." Information and views expressed in this document, including URL and other Internet Web site references, may change without notice. You bear the risk of using it. This document does not provide you with any legal rights to any intellectual property in any Microsoft product. You may copy and use this document for your internal, reference purposes. Licensed under [Creative Commons Attribution-Non Commercial-Share Alike 3.0 Unported](#)