Citation: Sandvig, Christian. (2006). Shaping Infrastructure and Innovation on the Internet: The End-to-End Network that Isn't. In: D. Guston & D. Sarewitz (eds.), Shaping Science and Technology Policy: The Next Generation of Research, pp. 234-255. Madison, Wis.: University of Wisconsin Press.

II

# Shaping Infrastructure and Innovation on the Internet

## *The End-to-End Network That Isn't*

CHRISTIAN SANDVIG

## Introduction

This chapter approaches the question of how we should best reason about the design of communication infrastructures by examining a particular debate about the Internet. Specifically at issue are the benefits of the Internet for innovation. Some argue that the Internet's gift is found in an obscure design feature called the "end-to-end argument," first elaborated by Saltzer, Reed, and Clark (1984). This design feature is a network engineering strategy that promotes "stupid" networks: the center lacks intelligence and performs only a few functions, while nodes at the edge of the network—the ends—build complex applications by employing the simple building blocks of the core. The Internet, with a smarter PC and a dumber router, manifests this design strategy, while the telephone network, with its dumber handset and smarter switch, does not. Proponents of the end-to-end argument hold that with the Internet's end-to-end design, experiments can be deployed from the edges (ends) by anyone at all. Success of the Internet

can be explained because experiments like the World Wide Web did just that.[1]

On the other side are commercial interests currently deploying intelligence inside the network's core. This logic speeds some traffic over others (caching), blocks traffic (firewalling, filtering), eavesdrops (snooping), and disguises some nodes as others (masquerading) (see David 2001). Some of these practices are so worrying for innovation—and for the freedom of users—that end-to-end proponents have asked the U.S. government to take action to preserve the Internet's "natural" form. Some have even argued that we might need a new Internet that retains an end-to-end design if the present one continues to erode.[2]

I take a third position. From the earliest organized communication systems, history teaches that networks tend to complexity at the core as more is asked of them. Contrary to the end-to-end argument, there is no reason to think that the Internet will reverse a three-millennium trend and evolve to be faster and more reliable than earlier communications systems while requiring fewer intermediaries and less intelligence at the center. While a more complex "middle" is already here, the important question has never been whether to preserve a simpler network structure but how and where new complexity is implemented. The key to innovation rests not on the degree of logic within intermediary nodes but in which nodes we trust. While Moors (2002) elaborates some of the technical implications of this position, this chapter addresses the implications for innovation and public policy.

I will address the future Internet by first recalling the oldest human communication networks—chiefly to remind us that although the structure of a communication network may have a technical veneer, it is a political bargain. Then, considering the Internet, I will unpack the end-to-end argument and suggest that: (a) it is not an organizing principle; (b) if it is a principle it is probably not true; and (c) even if it is true, it is probably not useful. The best outcome that normative claims premised on the end-to-end argument can offer us is to produce the right result for the wrong reasons, but we might be even better at promoting innovation if we act for the right reasons. Even worse, a dogmatic belief in end-to-end will simply retard the development of the Internet's infrastructure by limiting needed improvements in the "middle" or core. I will suggest that the right values to support are transparency and participation. I then conclude with the suggestion that underlying principles

that support innovation need to be addressed explicitly, not embedded tacitly in technical arguments.

## Internet Design Is an Old Problem

To tame the policy case of the present Internet, let us recast the present problems of high technology in the relatively sedate terms of technologies long dead; they are imperfect but at least relatively settled examples. Data networks have a long history. The first things we might qualify with the term "systems" of communication were made from human couriers. In this section, I will replace the esoteric language of computer network technology with examples of older systems that involved only humans in order to bring the social relationships implicit in network topography into sharper relief.

The first courier networks were point-to-point, with most of the intelligence about the network's condition located at the network's edge—there were no "courier network commissions" or planners of overall infrastructure. Messengers relied on their own knowledge of existing paths. Once an ancient messenger left for his destination, it was by no means assured that he would arrive. There were often financial and political benefits to intercepting messages, and "courier loss"—akin to Internet "packet loss"—could also be a simple case of highway robbery. As couriers are much harder and slower to replace than Internet packets, the solution was not to re-send lost messengers, but to change the network itself.

This complexity then distinguishes between two generations of ancient courier systems. The first generation system involved only couriers using whatever existing paths or routes were at hand. Second-generation courier systems had a greatly improved capability for signaling the state of the network, improved security, better reliability, and better performance. The first second-generation courier system known arose in China early in the Chou dynasty (approximately 1000 BCE). Typically, rather than relying on the paths that already existed, second-generation networks incorporated "sponsored" roads or "post" roads that were paid for through taxation. Money spent on marking and improving the quality of the road allowed the traffic to travel faster—examples abound from the Roman *viae militares* (the military road) to the Spanish *El Camino Real* (the king's highway). But often these were not

just upgrades to the same old infrastructure—beyond improvements to the old routes, the second-generation courier network added intelligence that was not at the network's edge.[3]

The Chinese system introduced repeaters: a system of "post-houses" allowed tired couriers to pass their messages to fresh, rested ones. In Chinese, Persian, Roman, and other civilizations, couriers on horseback later supplanted runner relays. Not only did the post-house or relay system increase the speed messages could travel, but post-houses were also used as a place to place guards to protect the integrity of the network from spies and robbers. They served a routing function by directing riders over alternative paths, and they maintained information on the quality of the routes. In some incarnations they billed for service, as toll stations. They filtered traffic—some post-house systems included armed guards and a capacity to inspect messengers. While Babylonian Royal Couriers were allowed to pass, the Bedouin raiders that hunted them most certainly were not (see Dvornik 1974).

One of the most ingenious additions to second-generation courier systems was the fire beacon. If the guard posts were numerous and located within sight distance, in clear weather the fire beacon could be used to send a prearranged message much faster than even a mounted courier, providing the network two modes of operation. These beacons could signal meta-information: in some systems they were used by the post-houses as a "trouble" signal warning of a problem with the network itself (a signal from the network's center to its edge, or as coordinating signal between different post-houses in the center). In special circumstances this faster mode could be quickly repurposed to carry simple information—often warning of invading armies. The integration of the beacon into courier networks (a protosemaphore) marks not just a general increase in the sophistication of the system but also the advent of segregating traffic into classes by priority, where each class has a different form and a different quality of service.[4]

## Policy Implications of Second-Generation Courier Networks

As it entailed investment in roads, the second-generation courier network required much greater standardization of traffic. Wheeled vehicles were nothing new—they predate the Chou second-generation courier network by about one thousand years—but when roads were improved, the road width had to be determined and future traffic had to conform.

Continuity was a decidedly local practice: at least eighty-four inches in width was required for a Roman road, but only fifty-five inches between shaped-stone wheel ruts sufficed for Greek sacred roads. These standards were then imposed on users by technology (the road) and also by law: the *raeda,* the fast freight cart used on Roman roads, was restricted by law to carrying 750 lbs. or less, for fear of damaging the road.[5]

Second-generation networks must have been much more effective than their predecessors and many orders of magnitude more expensive: guards, post-houses, and roads all cost money, but there were fewer dead messengers. The trend in the evolution of ancient courier networks is clear. As the demands on the networks increased, they became faster and more reliable in part by becoming more elaborate. As rulers added resources to the infrastructure, the communication network designers of the ancient world also added intelligence to the center of the network to manage and control it. They asserted this increasing control through a combination of force, law, and technology—in other words, the network's soldiers, rules, and form. As courier networks improved, the number of intermediaries must have increased at least slightly. No longer were couriers left on their own to wander. In the improved systems, the message might be normally expected to pass through many hands to be relayed, taxed, and inspected.

## The Internet and the Unexpected Reversal of History

In a hotly contested debate now occurring somewhere between network design and public policy, some argue that the optimal evolution of electronic networks will be just the opposite of these ancient courier networks—the Internet will evolve to be faster and more reliable than earlier electronic systems, and this can, in a reversal that is both paradoxical and exciting, require fewer intermediaries and less intelligence at the center of the network. In a sense, much of the early excitement about the Internet stems from this controversial design concept; it is the spring from which ideas about the Internet's "inherently decentralized nature" flow. In the remainder of this chapter I will unpack this idea, known as the "end-to-end argument," and then reassess it, employing the examples given above.[6] As additional examples will show, end-to-end has migrated from engineering circles to public policy circles. The focus of our interest here is the advocate's perception that the end-to-end argument in network design, usually evaluated by engineers on the

basis of technical efficacy, may also be normatively positive: that is, that the end-to-end argument forms a kind of philosophy of network topography. Normative benefits have been claimed for two reasons: First, under some conditions end-to-end networks can dramatically increase the number and diversity of groups that can participate in the design of network applications. Second, end-to-end networks can make it more difficult for unwanted third parties to control communication on the network. In this way, the end-to-end network has been hailed as inherently more democratic, its form producing user freedom.

However, I will argue that the current technical and political debates about end-to-end are misleading, for they cloak arguments about power with appeals to a single objective technical truth where none exists. Indeed, it is not end-to-end design per se that is normatively positive, but the transparency, openness, and participatory design consultation that have come to be associated with this model of network intelligence through history and tradition (Streeter 1999). Loading the end-to-end argument with these social goals, rather than addressing normative goals directly, is a dangerous and misguided strategy because it shifts policy discourse away from normative ends in favor of technical means that may not lead where we expect.

## What Is an End-to-End Network?

As is common in many areas of technological development, computer system and computer network development have exhibited since their inception a trend toward modularization. What was once a single piece of technology ("the computer") becomes an assemblage of different parts, some of which are standardized and produced by different parties. The innovations of the Internet reside in software, but modularization has still proceeded in a manner similar to the standardization of parts for a car. Today's programmers do not start by writing binary codes to control the hardware of a computer. Instead they assemble software by relying on a variety of standard components such as code libraries and features of the operating system. The modularization of software has enabled the growth of computing as a sociotechnical system: it facilitates innovation by allowing developers to incorporate standardized components in new and larger projects, increases reliability in software by distributing what is effectively pre-packaged expertise, reduces

software development time by changing the process from one of construction to one of assembly, and drives down software cost by allowing competition in development and provision of these modular subcomponents. The clearest expression of this modularization trend in the subdomain of computer *networking* has been the development of a standard model—called the seven-layer model—for representing communication between computers.[7] This model effectively divides the task of communicating into subtasks (layers), making the development of new network applications such as Web browsers and e-mail clients much simpler because the application programmer can leave most of the job of communicating to other subsystems. The entire means for sending messages does not require reinvention each time a new kind of message needs to be sent.

The seven-layer model is then useful for reasons of efficiency because it saves development cost. But as networking and the idea of network layers became widespread in the 1980s, so too did disputes about the technically correct level (that is, module or part) where any given functionality should be incorporated into a communication system. To translate this development to the ancient world, should the post-house count the couriers, or the king? Because different actors worked at different layers, these disputes could also take the form of the question: "Should this software company be responsible for a given task, or should that one?" In a now-classic paper in network engineering, Saltzer, Reed, and Clark (1984) elaborated a design strategy for computer systems called the "end-to-end argument," which take the position that:

> (1) If a particular function requires the participation of the endpoints of the system, it should not be implemented in any other location in the system.

This statement is the first of six principles that I extract from arguments made on behalf of end-to-end networks. It can be explained for the functional goal of reliability with a somewhat oversimplified scenario of mailing postcards:[8] If person A mails a very important postcard to person B, how does person A know that the postcard has arrived safely? One option would be to apply a variety of strategies at points in the postal system where loss of the postcard is likely to occur: the location of each item of mail in the system could be constantly tracked, duplicate postcards could be made at several points and delivered as insurance, or postcards could be fabricated out of a durable metal instead of

flimsy paper. The end-to-end argument, however, notes that while these strategies would increase the reliability of the postal system as a whole, they are expensive and might tend to slow the delivery of all mail. Indeed, regardless of any of these costly improvements to the postal system, person A could never be absolutely certain that person B did in fact receive a particular postcard. To attain peace of mind, person A would still have to ask person B if the postcard had arrived. It is likely more efficient, then, to skip the improvements to the postal system as a whole and for person A to plan on asking person B. As long as the use of the post is cheap and simple, if some of the postcards do not arrive, person B can ask for them to be sent again—by, for instance, using another postcard. In this scenario, person A and B are the "ends" and thus the repositories of all the intelligence necessary to confirm receipt of the message. No complexity was added to the middle (or core) of the postal network; instead it was the behavior of the "devices" connected to the edge of the network—the users—that changed to support the goal of reliability. In the eyes of the end-to-end proponents, simpler networks and smarter ends are always the answer.

## A Competing Strategy: The End-System Model

The early telephone network, in contrast, was initially developed with a network design philosophy that has been called "end-system" (Kruse, Yurcik, and Lessig 2001).[9] Most of the functionality of the network was located in the telephone switch, and the "end" (the telephone) was a fairly simple device that could do little except relay simple commands to the switch (off-hook, on-hook, 1234567890*#). To expand slightly on the postcard example above, we see that the end-to-end argument is clearly about who does what. If the communication problem is to insure that a postcard sent between person A and person B arrives, some end-to-end strategies might be that person A keeps very careful track of the postcard, A makes a duplicate of the postcard in case it is lost, or A chooses to make the important postcard out of a very durable material. In comparison, the end-system strategies would be that the post office implements a tracking system to keep very careful track of all postcards, the post office makes duplicates of all postcards when they are mailed, and the post office requires that all postcards be made more durable. And so in computer system design these questions about the location of functionality can seem like purely technical concerns, but the questions

drawn from this example are also clearly "who is in control?" and "who pays?"

End-to-end arguments have waxed and waned. When Internet pioneers first proposed end-to-end, it was controversial—end-system designs ruled the day in computer communication. In the related sphere of telecommunications, the dominant thinking was later called the "intelligent network" movement (Mansell 1994). In the late 1990s, backlash against this centralization led to a rediscovery (or independent discovery) of end-to-end in telecommunications under the catchphrase "the dumb network" or "the stupid network" (Isenberg 1997). Note that the categories of "end-to-end" and "end-system" are useful but not exhaustive. They are not a dichotomy, and indeed may not even be on a continuum. It is possible to characterize the structure of some networks as predominantly end-to-end or end-system, but other networks defy this characterization—they are confusing hybrids.[10]

Nevertheless, and despite the danger of over-generalization, table 11.1 attempts a rough categorization by way of example. As "end-to-end" and "end-system" networks are design strategies and ideal types, no network can clearly be labeled one or the other, but in this table I attempt to match the ideal types with a few examples that may come closest.

The earliest ad hoc courier networks were end-to-end in that the network was so simple that no complexity existed, except at the edge. All applications (diplomacy, commerce) were bargains between end-points. Although the earliest networks of paths were probably not, strictly speaking, "designed," still no intermediaries at all were planned. (Bandits are an example of an unplanned intermediary.) The earliest, nonswitched "party-line" telephone systems could also be conceptualized as end-to-end. Really they are broadcast networks, as all terminals received all transmissions, and there was no complexity at all in the middle of the network. This example brings us to a point of some confusion: the implied relation between end-to-end and switching. Note that the distinction between end-to-end and end-system networks is not related to switching. Switching is just one function of a network that might be implemented with an end-to-end or end-system strategy. For instance, the Internet is a packet-switched network that end-to-end proponents worry is becoming less end-to-end—but it will still be packet-switched. A nonswitched, that is, broadcast network could implement many functions in a complicated network core that are unrelated to switching. Last in the table is the network that gave birth to the concept

Table 11.1. Networks characterized by design strategy

| Ideal types: End-to-end networks | End-system networks |
|---|---|
| The earliest ad-hoc courier networks | The Chou Dynasty "post-house" courier network |
| Early nonswitched ("party-line") telephone networks | Early switched telephone networks |
| The Internet before widespread content caching | |

of end-to-end: the recent Internet before the advent of technologies that allegedly "break" end-to-end.

In the right column, the Chinese "post-house" system of the Chou dynasty could be considered end-system if we take some liberties with the analogy between courier systems and data networks. The post-house system also implemented routing at the core, but in addition it probably supported spying, taxation, and some other functions. It is a difficult example for this ideal type.[11] We also have the network where the concept "end-system" originated, telephone service before the advent of technologies that increase intelligence at the terminal device, or end (such as ISDN), and muddle the example. While the Internet's core handles routing—an example of some intelligence at the center—the Internet is still considered an end-to-end network because little other function was implemented in routers at the core.

This broad distinction between two ideal types in overall strategy for communication network design has consequences far removed from the practice of computing. The design of any technology that allows humans to communicate must have social and political consequences, but more specifically the perceived benefits of end-to-end design have been put forward in the case of the Internet in three areas: (1) user-driven innovation (2) protection from unwanted intermediaries, and (3) technical correctness.

## The Implications of End-to-End for User-Driven Innovation

Advocates of end-to-end point out that it reduces the complexity of the "core" network, and that the resulting generality of the network fosters innovation, because new and unanticipated services can employ the basic building blocks of a simpler network core (Blumenthal and Clark 2001). The innovation principle could be embodied succinctly as:

(2) The lowest layers of a system should provide the greatest flexibility possible, so as to permit applications that cannot be anticipated.

The Internet is a manifestation of this principle: it provides a fairly general set of facilities to allow the transfer of any data. A variety of different applications have emerged (e.g., remote login, electronic mail, the World Wide Web) that all share the same Internet. If certain functionality is required to allow only one of these applications to work, it is located in software (the electronic mail client, the Web browser) on computers attached to the network's "edge" (Blumenthal and Clark 2001, 92).

It is a key feature of end-to-end design that *users* of the network are able to create new applications—applications that eventually influence the broader technological system (cf. von Hippel 1988; Bar et al. 2000; Bar and Riis 2000).[12] This process of "user-driven innovation" occurs with other technologies, but because the control mechanisms for the Internet reside in software, user-driven innovation has a much more central role. Consider that the two most popular Internet applications were developed not by a centralized network authority or the owner of the network but by users: Ray Tomilson, who developed electronic mail for his group at BBN in 1973; Tim Berners-Lee, who developed the World Wide Web for his group at CERN in 1990.

By contrast, in an end-system model the device at the edge of the network—the telephone, for example—is simple and inexpensive to manufacture. The network equipment to which the device connects provides all of the intelligence and functionality of the system. In the early telephone network, adding new technology at the end was expressly forbidden—adding an answering machine or even a piece of cardboard was a violation of the system's design principles, and in some cases was illegal in the United States as late as 1976.[13] Both the design philosophy and the difficulty in configuring hardware act against user-driven innovation on the telephone network. Prior to electronic switching, an innovation such as three-way calling would have required a user to find a screwdriver and link three wires. Even since electronic switching and software control, if a user wishes to implement a new service such as voicemail on the telephone network, no facility exists to allow it. Telephone companies control the system, and the user may not tamper with it—or even discover how it works—without permission.

## End-to-End as Protection from Intermediaries

Those with technological determinist leanings have pointed out that the structure of the end-to-end Internet is itself a protection from anyone who would limit the freedom of communication. By this logic, adopting an end-to-end Internet design will produce freedom irrespective of legal or social arrangements. For much the same reasons that would-be innovators can deploy any new application that suits their fancy, would-be communicators do not need permission to speak and need not subject their speech to anyone's review. This argument says that if the network has no functionality to examine the messages it carries, communication is then more free.

## The Argument from Technical Correctness

Some have sought to offer end-to-end design principles as true in some objective sense. Principle 2 is less overtly concerned with objective correctness, but the next arguments considered are related to the idea that if functionality is removed from the network core, the core itself likely becomes cheaper, faster, and easier to administer. They argue that a simpler core is necessarily more transparent and easier to model, and that acceptance of this argument is a step toward a more rational, scientific, and rule-based network engineering (for an account and critique of this specific point, see Moors 2002). These "correctness" arguments have been made and remade in the landmark papers on end-to-end, in fora like the end-to-end mailing list and, one imagines, in engineering meetings around the globe with increasing frequency over the last twenty years. It has been claimed that end-to-end is technically correct because:

> (3) Any function implemented in the core network may be redundant because this functionality has already been implemented at the end-point.
> (4) Any function implemented in the core network may be redundant because some applications will never need it.

Let me return again to the postcard example to explain these claims of end-to-end proponents. To illustrate principle 3, consider that if both the end-to-end and the end-system strategies in table 11.1 are implemented,

there will be a significant duplication of effort. Furthermore, to illustrate point 4, imagine person C who does not need to verify if her postcards are delivered or not (or does not wish to pay extra for it). If person C pays for the postal system (through stamps or taxes), part of her contribution to the postal system through taxes will go to support improvements she does not need or want. The network will be more expensive to operate, with no benefit to her. A stronger form of statement 4 is that no function should be implemented in the network unless all clients of the network (or that network layer) will need it, because:

> (5) The end-point tends to have more information about what it needs than the network, and
> (6) Any function implemented in the core network adds cost and complexity that is borne by all network users, even if they do not use the function.

While principle 5 seems to be an argument for end-to-end design, it presupposes an end-to-end network where there is no central authority dictating what applications should be used, and where the technology and form of communication remain unsettled. If the Internet will continue to be a place where new applications arise continually (e.g., peer-to-peer file sharing in the form of Napster) there is some merit to this point. Statement 6 is merely a rearticulation of the conclusions arrived at earlier with the postcard example.

## The Perceived Challenge to End-to-End

Attempts to promote new Internet applications have proponents of end-to-end design feeling as if they were under siege. Proponents have warned that developments in software, policy, and use are "compromising the Internet's original design principles" (Blumenthal and Clark 2001). Blumenthal and Clark note four examples of emerging requirements for Internet applications that challenge end-to-end design principles: (1) the need to manage untrustworthy end-points (2) demands for better throughput required by streaming audio and video (3) differentiation of service between competing Internet Service Providers (ISPs), and (4) the rise of increasing third-party involvement in communication. That is, each of these developments might imply a need to add intelligence to the network, for example (1) the "firewall" to block "hostile"

network traffic (2) the addition of proprietary distributed cache systems for multimedia content (e.g., Akamai) (3) the distinction between different kinds of content by the ISP to provide different levels of service quality, and (4) the use of filters to block unwanted or illegal traffic, or the use of traffic analyzers to eavesdrop on suspect traffic.

The perceived danger of these changes is that they each constitute new intermediary points in the path of network traffic where some third party may exert control. Those concerned about the censorship of content or the leveraging of the control of the Internet's wires into the control of its content have pointed out that "end-to-end was initially chosen as a technical principle. But it didn't take long before another aspect of end-to-end became obvious: It enforced a kind of competitive neutrality. The network did not discriminate against new applications or content because it was incapable of doing so" (Lessig 2000). Indeed, one of the original authors of the seminal end-to-end paper equates end-to-end with "the default situation [where] a new service among willing endpoints does not require permission for deployment." Abrogation of end-to-end design principles leads to the case where "new chokepoints are being deployed so that anything new not explicitly permitted in advance is systematically blocked" (Reed 2000, 4).

The strength of the arguments made to ensure future end-to-end design relies on the presumptions that end-to-end is the current design scheme, and that the current design scheme has been an effective one. That is, if it ain't broke, don't fix it. As Lemley and Lessig (2000, 4) state, "We do not yet know enough about the relationship between these architectural principles and the innovation of the Internet. But we should know enough to be skeptical of changes in its design. The strong presumption should be in favor of preserving the architectural features that have produced this extraordinary innovation." These appeals have been made in policy fora and seek to constrain those who would depart from end-to-end design principles. These threatening parties are private actors, such as ISPs and telecommunications companies, free to attach their new software and equipment to the Internet as they see fit. In other words, end-to-end proponents wish to make the case that while the Internet has a "fundamentally" decentralized and distributed nature, it now requires policy action to prevent private actors from departing from its technical design principles.

The most visible recent manifestation of this conflict has been the U.S. cable industry "open access" debate. Cable providers in the U.S.

have deployed extensive intelligence in the network's middle in the form of a technology called a "caching gateway" at the junction of their subscriber network and the slower Internet. Because cable providers already own government-sanctioned monopoly franchises, if they are allowed to require that users of broadband cable modem service also use an ISP that they own, it is likely that they will be tempted to leverage their monopoly power into control of Internet content. This leveraging could be accomplished by hosting "strategic partners" or in-house content on these caching gateways that they alone control and that are close to cable modem subscribers. This arrangement would provide subscribers with faster access to content that generates profits for the cable company, and indeed also provides the cable company with a temptation to slow traffic from competing entities. The user in this scenario would never know why some parts of the Internet were faster and some slower. The U.S. Department of Justice, the Federal Trade Commission, and the Federal Communications Commission were lobbied with this rationale to require open access to competing ISPs in merger proceedings involving large cable providers (Bar et al. 2000).

## End-to-End: Already Over or Never Was?

In this section I critique end-to-end approaches in order to show that the end-to-end debate is not one of technical correctness but of sociopolitical control. First, I should point out that what has been presented in the section above is a classic case of a technology in the early period of its development, when its inner workings are subject to interpretive flexibility (Pinch and Bijker 1987). Several relevant social groups have been identified in this controversy, for example, private firms seeking to profit from the Internet in some way, and the "old school" of Internet pioneers, designers, and computer scientists. The Internet is a technology that they each seek to shape, and in order to assert control over it they are engaging in a definitional controversy about what features are essential to the object "Internet." They attempt to control the emerging technology by painting the encroachments of other groups as antithetical to the natural form of the thing that we call "Internet."

As the authors of the original end-to-end paper note, end-to-end is an "argument" rather than a rule, law, or even principle. Determining what is meant by "ends" in any given engineering problem is extremely

difficult, and the resulting debates are open to much interpretation. Even the strongest case made for end-to-end, the case for the promotion of innovation, is problematic. The end-to-end approach only facilitates innovation of the kind where the new application services can be built with the low-level building blocks provided by the network. Since network engineers do not agree on a kind of "periodic table" of low-level network building blocks with which every possible service can be built, the end-to-end goal of a simple and easy-to-build-from network is actually an easy-to-build-from network where it is easy to build things that are made from the blocks or functions that are present. If the service you want to build cannot be built from these functions, an end-to-end design strategy will not make your new idea easier to build.

For instance, those who wish to introduce robust multimedia streaming services object that it is not possible to build quality of service guarantees when starting from the building blocks that are available. To work at all, the development of streaming for multimedia content seems to require changes to the core functionality of the network. It is hopelessly impractical to treat broadcasting as a point-to-point operation in the present Internet. With any large number of users wanting to view a broadcast stream simultaneously, the load on the provider of the stream quickly becomes unmanageable, and the network near the provider is subject to congestion because the provider must send multiple identical copies of the stream, one for each user that desires it—an ugly and inefficient approach. The "stupid" network does not notice that each stream is the same thing and cheerfully sends a thousand copies when one would be best.

The cooperative, open approach to redesigning the Internet's core to support this semibroadcasting of multimedia content is termed "multicasting," and it involves modifying the Internet's basic protocols to reduce the duplicate transmission of multiple streams when one stream would do. End-to-end proponents have described these core network modifications as a necessary departure from an end-to-end strategy for reasons of performance.

However, providers of privately distributed stream-caching services now offer reliable broadcasting on the Internet without modification of existing protocols or the network's core. These providers, Akamai and formerly Inktomi, for example, locate servers in many data centers around the world. The content stream is sent first to these data centers, where it is then duplicated for users that are nearby. In other words,

users are directed to obtain the content from the point that is nearest them. This approach has been decried by end-to-end proponents as a violation of end-to-end.

Clearly the approach used by Akamai reduces the transparency of the network. Akamai is a private company, and it has introduced a proprietary facility for providing content that is available only to its subscribers. In contrast, if multicasting were fully realized within the Internet's protocols, it would provide a facility inside the network for providing such content that could be used by anyone. In effect, a key difference between these two approaches is cost distribution. Modifying the core network, as Akamai does not do, distributes the cost of broadcasting Internet content across all users: those who want the content, those who provide it, and those who never use it. The Akamai approach requires that multimedia content providers and providers of popular (nonmultimedia) content pay for this infrastructure as an added service, in advance. The decision of where to locate the functionality determines how Internet broadcasting is funded and who has access to it.

## Conclusion

In considering policy decisions related to the design of the Internet, it is easy to despair when a decision must be made without much precedent. But the history of communication networks abounds with useful comparisons in the distribution of intelligence within networks. Indeed, the present debate can be usefully informed by the history of courier networks and the difficulties of asserting control through network protocols. The word "Internet" evolved from what ARPANET protocol designers called "the Inter-network problem," that is, the between-network problem. "Networks represent administrative boundaries of control, and it was an ambition of [the ARPANET] to come to grips with the problem of integrating a number of separately administrated entities into a common utility" (Clark 1988, 107). Coordination of the second-generation courier network of the Chou dynasty, then, needs to be extended further still, to ask what happens when the couriers reach the edge of a kingdom and must pass onto courier networks controlled by other kings. To borrow the language of computer networking, the network designers of ancient China did not consider technical means for ensuring the integrity of their communication across these administrative

boundaries of control, because any protocol they might devise would depend entirely upon the cooperation of rulers of neighboring kingdoms. If it was in the strategic interests of other rulers to intercept or otherwise compromise messengers, the problem is not a technical one but one of politics. The only way to make sure that adjacent interconnected networks behave the way you want them to is to assert control over them in some fashion.

The use of end-to-end as a design principle has the effect of pushing intelligence to the borders of the network, creating what end-to-end advocates argue are "application insensitive" networks that are optimized to deliver few "low-level" services at the core. They also claim that these networks offer not only building blocks upon which many types of use can be built but also an environment where there need be few intermediaries. I hope to have demonstrated that this "low-level" of service is better described as one form of service among many possibilities, for the Internet is application-insensitive as long as the application is similar to something its first designers envisioned. The Internet has always had intermediaries, but end-to-end proponents were happiest with the intermediaries they knew—service providers who initially were academic computer scientists and then universities. The end-to-end argument is an effort to stop the new intermediaries by arguing that they are technically incorrect—a definitional debate about the form of the technological artifact "Internet," which has not yet stabilized.

I do not intend to underestimate engineers. My argument does not imply that those who advance end-to-end arguments as technical solutions to their political problems are in any way simpleminded. Many of the participants in these debates are well aware of the dangers of advancing normative principles as technical principles, or conflating the technical and the normative. But the technical argument is alluring because it offers the promise of objective correctness to trump messy compromises.[14] Instead, I suggest that these clever engineers are themselves underestimating policymakers and corporate opponents. To stand on expertise instead of principle will work only if no other experts contradict you. When those who wish to modify the network in "undesirable" ways do not like your engineering, they will simply buy new and better engineers. The use of technical arguments as proxies for normative arguments produces a strange debate. If one argues that "the end-to-end principle renders the Internet an innovation commons" (Lessig 2001, 40), one then cedes the debate to whichever specialists are successful at

defining the historical, or true Internet. In arguments from tradition, those who can define the past get to define the future.

It is seductively easy to conceptualize technical, social, and legal mechanisms of control as different sorts of levers one can pull to steer the Internet (Clark et al. 2002). Writing in the end-to-end debate has pointed out the prevalence of technical control, and it has also called for more social and even legal control as a better "balance." In fact, what seems like "technical" control is nothing new. The technical, social, and legal always interpenetrate, and there is no purely technical way to guarantee that the Internet will evolve or be used in one way and not another without some broader assertion of control, by owners or by governments.

The process of network design should continue to include considerations of transparency, participation, and flexibility, but these should be explicit goals and not pursued under the rubric of technical correctness or the end-to-end argument. Furthermore, the legitimate public policy role for governments lies not in protecting the Internet against those who would "break" it. Such a policy would merely grant authority to whoever is designated to interpret the Internet's fundamental nature and write its history. Reflecting on the Internet's boon to innovation provides a logical rationale for regulating transparency and participation. This role is not a new one for government, even with respect to the Internet.

## NOTES

The author would like to thank Helen Nissenbaum, Stephen Barley, Paul David, Ian "Gus" Hosein, William Drake, and Dieter Zinnbauer for their helpful suggestions. This research was kindly supported by the Markle Foundation Information Policy Fellowship at the Programme in Comparative Media Law and Policy at Oxford University, and by a Visiting Research Fellowship at the Oxford Internet Institute. An earlier version of this chapter that did not contain the arguments related to innovation was presented at "Computer Ethics: Philosophical Enquiries," 15 December 2001, Lancaster, UK.

1. The most visible popularizer of this notion may be Lessig (2001), the most precise Blumenthal and Clark (2001).

2. This intriguing suggestion was made by David (2001) and also deemed impractical by him.

3. The historical details in this introduction are taken from Holtzmann and Pehrson (1995).

4. The simple beacon of light (fire) or smoke may predate the organized courier network, but the point here is the integration of the two.

5. The details in this paragraph are taken from Lay (1992) and Forbes (1954).

6. In telecommunications engineering, the notion of a single carrier owning all segments of a connection between two parties is also called "end-to-end," but this use of the term is not relevant to this paper. This paper's use of "end-to-end" refers to the definition used in computer networking from the 1970s forward.

7. The seven-layer model was developed as part of the Open Systems Interconnect (OSI) initiative of the International Organization for Standardization (ISO).

8. The original example given in the article was the problem of "careful file transfer" (p. 510).

9. Meaning that as an approach to designing the system, the design of the end is oriented to communication with the system (in this case, the switch) and not another endpoint. To simplify the term it might be easier to conceptualize it as "from end to system."

10. For instance, Integrated Services Digital Network (ISDN), a way for computers to communicate via telephone deployed in the 1980s, was an attempt to move some intelligence from the center of the (end-system) telephone network to the edge, and to implement some features in smarter terminals (called ISDN terminal adapters) than the traditionally dumb telephone.

11. Stations in military courier networks like the post-house system were usually also distribution points for news and gossip and accepted nonmilitary traffic when extra capacity was available. We can imagine, if we adopt today's terms, a richer topography with nonuniform nodes, some broadcasting, some caching, and some filtering. All of this was, if not planned, at least accepted as normal.

12. Note that "users" may not necessarily mean novices—some users have considerable skill with the technology they are using.

13. See, for instance, a review of the Hush-a-Phone and the Carterfone (Neuman, McKnight, and Solomon 1998, 176–78).

14. In this section I take issue with those who cover for normative goals with technical conclusions, but it is doubtful that the distinction between "technical" and "normative" is ever very useful. Every technological decision includes normative assumptions, even if these are so generally accepted or unexamined that they are ignored.

## REFERENCES

Bar, F., and A. Munk Riis. 2000. Tapping user-driven innovation: A new rationale for universal service. *Information Society* 16 (2): 99–108.

Bar, F., S. Cohen, P. Cowhey, J. B. DeLong, M. Kleeman, and J. Zysman. 2000. Access and innovation policy for the third-generation Internet. *Telecommunications Policy* 24 (6/7): 489–518.

Blumenthal, M. S., and D. D. Clark. 2001. Rethinking the design of the internet: The end-to-end arguments vs. the brave new world. In *Communications Policy in Transition: The Internet and Beyond,* ed. B. M. Compaine and S. Greenstein. Cambridge, MA: MIT Press.

Clark, D. D. 1988. The design philosophy of the DARPA Internet Protocols. *Computer Communication Review* 18 (4): 106–14.

Clark, D. D., J. Wroclawski, K. R. Sollins, and R. Braden. 2002. Tussle in cyberspace: Defining tomorrow's internet. Paper presented at the Annual Meeting of the ACM Special Internet Group on Data Communications (SIGCOMM), in Pittsburgh, PA. 19 August.

David, P. A. 2001. The evolving accidental information super-highway. *Oxford Review of Economic Policy* 17 (2): 159–87.

Dvornik, F. 1974. *Origins of Intelligence Services: The Ancient Near East, Persia, Greece, Rome, Byzantium, the Arab Muslim Empires, the Mongol Empire, China, Muscovy.* New Brunswick, NJ: Rutgers Univ. Press.

Forbes, R. J. 1954. *Roads to c. 1900.* Vol. 4 of *A History of Technology,* ed. C. Singer. Oxford: Clarendon Press.

Holzmann, G. J., and B. Pehrson. 1995. *The Early History of Data Networks.* Los Alamitos, CA: IEEE Computer Society Press.

Isenberg, D. S. 1997. The rise of the stupid network. *Computer Telephony* (August): 16–26.

Kruse, H., W. Yurcik, and L. Lessig. 2001. The InterNAT: Policy implications of the Internet architecture debate. In *Communications Policy in Transition: The Internet and Beyond,* ed. B. M. Compaine and S. Greenstein. Cambridge, MA: MIT Press.

Lay, M. G. 1992. *Ways of the World: A History of the World's Roads and the Vehicles That Used Them.* New Brunswick, NJ: Rutgers Univ. Press.

Lemley, M. A., and L. Lessig. 2001. The end of end-to-end: Preserving the architecture of the Internet in the broadband era. *UCLA Law Review* 48: 925–72.

Lessig, L. 2000. Innovation, regulation, and the Internet. *American Prospect* 11 (27 March). Available at http://www.prospect.org/print/V11/10/lessig-1.html.

———. 2001. *The Future of Ideas: The Fate of the Commons in a Connected World.* New York: Random House.

Mansell, R. 1994. *The New Telecommunications: A Political Economy of Network Evolution.* Thousand Oaks, CA: Sage.

Moors, T. 2002. A critical review of "End-to-End Arguments in System Design." In *Proceedings of the IEEE International Conference on Communications (ICC).* Vol. 5, , 1214–19. New York: IEEE.

Neuman, R. W., L. McKnight, and R. J. Solomon. 1998. *The Gordian Knot: Political Gridlock on the Information Highway.* Cambridge, MA: MIT Press.

Pinch, T. J., and W. E. Bijker. 1987. The social construction of facts and artifacts: Or, how the sociology of science and the sociology of technology might benefit each other. In *The Social Construction of Technological Systems: New Directions in the Sociology and History of Technology,* ed. W. E. Bijker, T. P. Hughes, and T. Pinch. Cambridge, MA: MIT Press.

Reed, D. P. 2000. *The End of the End-to-End Argument* [online]. April 2000 [cited 17 September 2004]. Available at http://www.reed.com/Papers/endofendtoend.html.

Saltzer, J. H., D. P. Reed, and D. D. Clark. (1984). End-to-end arguments in system design. *ACM Transactions on Computer Systems* 2 (4): 277–88.

Streeter, T. 1999. "That deep romantic chasm": Libertarianism, Neoliberalism, and the computer culture. In *Communication, Citizenship, and Social Policy: Re-Thinking the Limits of the Welfare State,* ed. A. Calabrese and J. C. Burgelman. New York: Rowman & Littlefield.

von Hippel, E. 1988. *The Sources of Innovation.* Oxford: Oxford Univ. Press.

# Shaping Science and Technology Policy

## *The Next Generation of Research*

Edited by
DAVID H. GUSTON
and
DANIEL SAREWITZ

THE UNIVERSITY OF WISCONSIN PRESS

The University of Wisconsin Press
1930 Monroe Street
Madison, Wisconsin 53711

www.wisc.edu/wisconsinpress/

3 Henrietta Street
London WC2E 8LU, England



1 3 5 4 2

Printed in the United States of America

Library of Congress Cataloging-in-Publication Data
Shaping science and technology policy: the next generation of research /
edited by David H. Guston and Daniel Sarewitz.
p. cm.—(Science and technology in society)
Includes bibliographical references and index.
ISBN 0-299-21910-0 (cloth: alk. paper)
ISBN 0-299-21914-3 (pbk.: alk. paper)
1. Science and state—Decision making.
2. Technology and state—Decision making.
3. Research—International cooperation.
4. Science and state—Citizen participation.
5. Technology and state—Citizen participation.
I. Guston, David H. II. Sarewitz, Daniel R. III. Series.
Q125.S5164 2006
338.9′26—dc22 2006008594