The Concept of Embedded Values and the Example of Internet Security

Aimee van Wynsberghe, Giovane C. M. Moura*
University of Twente
Centre for Telematics and Information Technology (CTIT)
*Faculty of Electrical Engineering, Mathematics, and Computer Science
*Design and Analysis of Communications Systems (DACS)
Enschede, The Netherlands
Email: {a.l.vanwynsberghe, g.c.m.moura}@utwente.nl

Abstract

Many current technological devices used in our everyday lives confront us with a host of new ethical issues to be addressed. Facebook, Twitter, or smart phones are all examples of technologies used quite pervasively which call into question culturally significant values like privacy, among others. The embedded values concept presents the compelling idea that engineers, scientists and designers can create technologies which intentionally enhance cultural and societal values while at the same time minimizing threats to other values. Although the embedded values concept (and the resulting design theories that follow) is of great utility, it remains unclear how to utilize this concept in practice. Added to this is the difficulty of utilizing this concept when engaged in fundamental research or experiments rather than in the creation of a commercial product. This paper presents a novel approach for collaboration between an ethicist and a computer engineering PhD researcher working on the Internet Bad Neighborhoods concept for spam filtering. The results proved beneficial in terms of both the utility of the embedded values concept as well as a strengthening of the engineering PhD researcher's work.

1 Introduction

Many current technological devices used in our everyday lives confront us with a host of new ethical issues to be addressed. Social networking sites, like Facebook and Twitter, are examples of technologies used on a daily basis that encourage us to reflect on our conception of values like privacy, among others. The embedded values concept in computer ethics [1] has drawn the attention of ethicists and designers alike to the relationship between the design of technologies and the resulting ethics when the technology is used[2, 3, 4].

We take the main argument of the embedded values concept as our starting point but we claim that the predominant difficulty facing its uptake lies in exactly how this can be done in practice. For example, how exactly does one conceptualize and translate the values relevant to the creation of the technology in question in order for them to be embedded?

The purpose of this paper is to address the *concept of embedded values by outlining a methodology in which ethicists and computer engineers can collaborate in order to conceptualize values* when research is still in the fundamental stages. To accomplish this, we will present the collaboration between an ethics advisor at a technical research institute (van Wynsberghe) with a then Ph.D. candidate (Moura) at the Computer Science department, working on a project investigating the concept of Internet bad neighborhoods to improve network security [5]. Briefly, the main idea behind the Internet bad neighborhood concept is that Internet traffic can be filtered based not only on the source's behavior, but on behavior of its surrounding neighborhood. For example, an e-mail filter can classify a message from an unknown source solely based on the sources neighbor's behaviors.

In an interdisciplinary investigation, the authors worked together to scrutinize the engineer's intended values in the Internet neighborhood concept. Our steps were: making intended values explicit, scrutinizing these values, and balancing values with context in mind. The approach presented here is a novel approach to the research and design of ICT systems, and it proved to be beneficial in both the justification of the concept of embedded values, and value analysis in general, as well as strengthening the integrity of the work of the engineer.

2 The embedded values concept

The relationship between the use of technologies and the idea that values are embedded in technologies has only recently gained attention. It presents a compelling idea: technologies can embed values such that when they are used certain values come into existence [1]. For example, when Facebook users are using the social networking site, the value of privacy comes into existence. This does not mean it is the conception of privacy that all citizens would agree with but a certain conception of the value of privacy is made real through the use of the platform.

From this comes the idea that if we agree with the claim made for embedded values, then we can design technologies to intentionally promote certain values and minimize threats to other values. This is a very powerful idea and is the driving force behind Value-Sensitive Design (VSD) [6] and other value conscious design methods that aim to incorporate value analysis into the design process. Taking the Facebook example into consideration again, one might claim that a re-conceptualization of Facebook be done for future applications and/or uses to minimize privacy threats of users.

Although the embedded values concept seems to be a promising concept for addressing the relationship between the use of technologies and the realization of values, it still faces a host of criticisms[7, 8, 9]. Of particular interest for this paper is the latest criticism addressing the disconnect between the intended values of engineers and the values realized once the artifact is used in context [7, 8]. In other words, even when an engineer may intend a value in the design of a technology this does not guarantee that the value will be realized in practice.

3 Bridging the gap between Ethics and ICT research and design

The main question of interest for this work is how to utilize the embedded values concept when engaged in fundamental or experimental work rather than in a commercial

or industrial setting.

The first step to embedding values in technologies is making those intended values explicit. With this in mind value trade-offs must also be brought to the fore. A value trade-off exists when the presence of one value minimizes the presence of another. This forces the engineer to question whether or not such a trade-off can be justified.

The methods used here to bridge the gap in ethics and ICT research and design are based on the work of an ethics adviser (van Wynsberghe), for a technical research institute, CTIT (Center for Telematics and Information Technology). Her role as ethics advisor has her committed to the incorporation of ethics into ICT research and design. This may happen in a variety of ways and is dependent on: the kind of research (fundamental, experimental or the development of a product), the stage of development (i.e., the design process), the type of product, and the design team goals/wishes. That being said, each collaboration centers on a value analysis involving: making intended values explicit, questioning these values, and balancing values with context in mind (examining value trade-offs). For the work presented in this paper, the value analysis is done using the concept and research of Internet bad neighborhoods and thus focuses on fundamental research (vs. the creation of a product) at the final stages in the development stages.

4 The Internet bad neighborhood concept

The Internet is currently so important for the functioning of our society that it is actually considered part of the critical infrastructure of many countries [10]. Such dependence has made the Internet very attractive for criminal organizations, nation states, and activists as a medium in which crimes, cyber war, and protests can be carried out. One example is spamming, in which cyber gangs generate massive amounts of spam messages for various purposes (selling counterfeit drugs, spreading viruses, etc.). It is estimated that 84-90% of all e-mail messages are spam nowadays [11, 12], which ultimately leads to estimated losses of $10 billion to $87 billion yearly [13].

Behind these attacks, we find a large number of compromised computers, which are typically computers in homes, schools, and businesses that have been "hijacked" and carry out their malicious activities. This is done in a stealthy way so the human user behind the desk does not notice their computer has been hijacked. Even though these computers are distributed all over the world, they are, in fact, concentrated in certain networks instead of being evenly distributed. Figure 1 shows the number of spamming hosts per network in an experiment from [5] (/8 netlblock, in IP addressing scheme[1]). As can be seen, some networks (or netblocks, x axis) exhibit higher numbers of spamming hosts (y axis) than others. This resembles the distribution of crimes in the real world. For example, Figure 2 shows that homicide locations are more concentrated in certain neighborhoods in New York City than others.

With this in mind, if the New York Police Department (NYPD) wants to reduce crime more efficiently, a starting point would be to improve police coverage where the homicides are concentrated – for example, in neighborhoods like Brooklyn or the Bronx. The same principle applies to bad neighborhoods on the Internet. If network security engineers (analogous to the NYPD) want to reduce the incidence of attacks on the Internet, they should start by tackling networks where attacks are more frequently originated. Based on this, the Internet the bad neighborhood concept emerged [15].

[1] Please refer to [14] for a brief description on the subject matter.

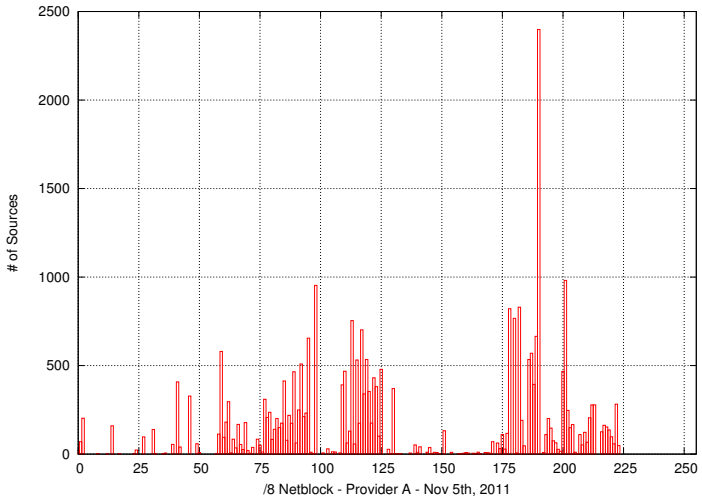

Figure 1: Number of spam Sources per /8 netblock

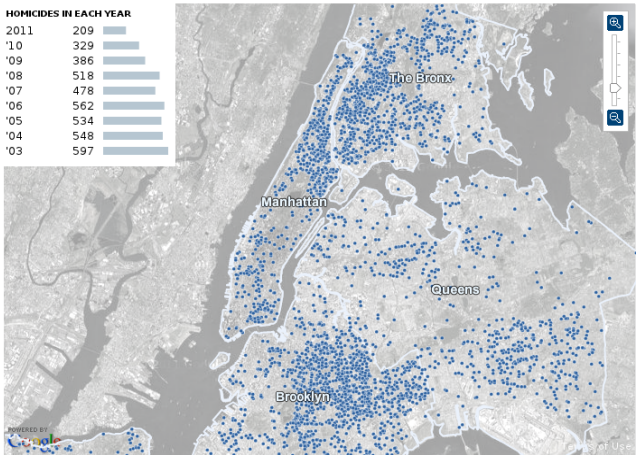

Figure 2: New York City Homicide Map (2003-2011) – Source: *New York Police Department*

5 Value analysis of the Internet bad neighborhood concept

To begin the ethical evaluation of the Internet bad neighborhood concept, three ethical dimensions were uncovered to be addressed: 1. the labeling of malicious hosts, 2. the labeling of bad neighborhoods and, 3. the deployment of the concept. The first two categories pertain to the concept of Internet bad neighborhoods in general, independent of any deployment of the concept, while the third category required a specific context or deployment in mind in order to discuss.

A value analysis was then carried out for each of these three dimensions. Each of the categories presented differing ethical considerations; however, what all three did have in common was that they relied on a comparison between real world bad neighborhoods and Internet bad neighborhoods. In other words, it became clear that

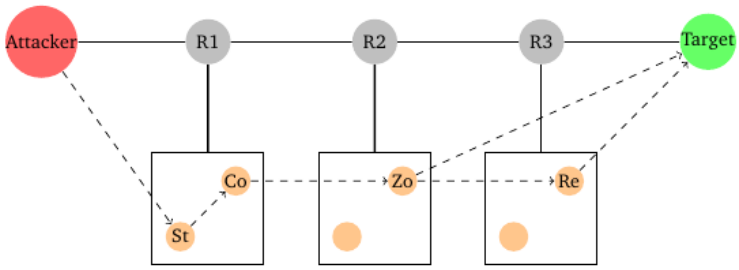

Figure 3: Attribution Problem (adapted from Wheeler and Larsen [17])

the same manner in which individuals may suffer from bias or discrimination in the real world, may be manifest on the Internet through the use of the concept.

5.1 The labeling of malicious hosts

The labeling of malicious hosts refers to how hosts are "flagged" as wrongdoers. In the real world, lists of offenders are kept by police departments. On the Internet, however, that is not a trivial task: in fact, most of attackers hijack computers online, so they can pass by their victims unnoticed when carrying out attacks. To illustrate this, see Figure 3, in which the real attacker (`Attacker`) attacks the `Target` computer employing a series of intermediate computers (`St, Co, Zo, and Re`) before reaching its `target`. If the attack is noticed, it will be perceived by the `Target` as if it was originated from the last host in the sequence (`Zo or Re`, depending on the chosen path). In this case, an innocent computer system may be blacklisted by the `Target`, while the original attacker might be thousands of miles away and escape unidentified[2]. As a consequence, in this case, the legitimate user in `Zo` might not be able to send legitimate e-mail to the mail server at the `Target`. This is known in the security community as the attribution problem [17].

For this dimension, the intended value of the engineer was Internet security. The main ethical concern for this dimension, however, had to do with the potential for mistaken bias and prejudice, ultimately resulting in a minimization of the value of fairness. Mistaken bias and prejudice had to do with what would happen to users who had wrongfully been labeled as malicious hosts (`Zo or Re`) but who were innocent bystanders – the attribution problem [17]. This is an example of a value trade-off – the value of Internet security comes at the price of the value of fairness. To account for this value trade-off, Moura relied on a utilitarian approach in ethics and claimed that in order to provide the value of security to the greatest number of Internet users this compromise would have to be made.

5.2 The labeling of bad neighborhoods

The labeling of bad neighborhoods refers to how hosts are clustered to form a bad neighborhood. To illustrate this, see Figure 4, in which the target A is attacked by hosts 3 and 4 (left side). In a traditional network security approach, both hosts could

[2]To make the attribution problem even more complex, the attacker may benefit from other network features. Since the IP source address of the attackers is not used in the routing process, it may be easily forged, which is commonly known as IP spoofing [16]. Other techniques can also be employed; for a more detailed view on the matter, please refer to the work of Wheeler and Larsen [17].

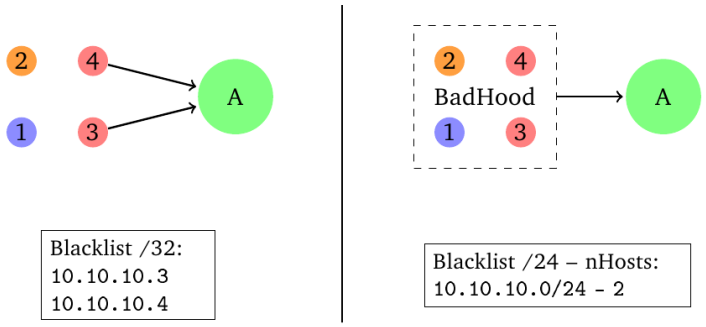

Figure 4: Aggregating Malicious Hosts into BadHoods

potentially be blacklisted, and any e-mail from both in the future would be potentially rejected. However, by employing the bad neighborhood concept (right side of the same figure), not only hosts 3 and 4 are blacklisted, but also their neighbors (1 and 2), even though they have not previously sent any spam to A. In fact, it is like the whole neighborhood has carried out the attack instead of the individual hosts.

For this ethical dimension, it was revealed that the intended value of the engineer was once again Internet security. The ethicist, at this time, was concerned with the reliability of the clustering of Internet bad neighborhoods and once again the potential for mistaken bias and prejudice. Now, however, the mistaken bias and prejudice was made worse given that an address could have mistakenly been labeled a malicious host (i.e. the attribution problem [17]) and then all the neighboring addresses of the e alleged "malicious" host were aggregated and labeled as a bad neighborhood. The question of the ethicist at this time was: would it not be fairer to the public if IP addresses were *randomly* assigned to a bad neighborhood? Would this not prevent a minimization of fairness to labeled hosts? This question invites a discussion of what will happen once the concept has been deployed. For this Moura and van Wynsberghe speculated on potential future uses, users and outcomes as discussed below.

5.3 The deployment of the concept

Of particular interest here was that the computer engineer was conducting fundamental research on the topic versus the creation of a commercial product. The deployment of Bad neighborhood-based network filtering technologies would be the responsibility of the software vendor. As discussed by the engineer[5], the bad neighborhood approach is the "final" solution for Internet security problems, and should be employed as a compliment to existing solutions. Consequently, envisioning the intended uses and users was not the focus of the study; however, in order to understand the value trade-offs and the intended values in general, a use context needed to be envisioned and explored.

Added to the overarching intended value of Internet security the engineer intended that the model be reliable, fair and effective. At this stage it is important to scrutinize the intended values to ensure they are in fact desirable for society. In general, these are all values considered relevant and desirable to society at large but the question of greatest ethical importance for this paper is whether or not they would in fact be realized and what other values were being traded-off to ensure these.

In making clear the ethical issues related to the first two dimensions and how they

could become problematic once the technology was deployed the value trade-off became clearer. In other words, if a company were to use this model and IP addresses were labeled malicious which then meant that the neighboring addresses were also labeled malicious, what would happen to the emails sent from these addresses? Would they bounce back to the sender or be thrown out without any indication of doing so? What would happen if home systems that were also connected to the Internet (e.g, fridges, satellite receivers, ADSL routers,etc.) were labeled as a malicious host, when clearly it was not the original attacker, and all systems in the Internet neighborhood were compromised without an indication of this being possible (meaning, no email could be sent or passed on from the fridge). Where would responsibility fall when the labeling had happened like this and still we do not know if the idea to cluster based on address was in fact reliable?

Making the intended uses explicit and taking them into consideration it became clear that the intended values of reliability, fairness and effectiveness may not be realized once the concept was employed. Specifically, how could the claim be made that the concept was reliable and effective if it might be just as effective to randomly assign IP addresses to a bad neighborhood? Added to this, how could it be claimed that the concept was fair to citizens if it could not be guaranteed that it was reliable and effective? This conflict in values had not occurred to the engineer before this time.

6 Balancing value trade-offs

The fact that the intended values of reliability, effectiveness and fairness may not be realized once the concept was deployed presented the engineer with a value conflict. To account for this, the researcher conducted a series of additional experiments. These additional tests were aimed at supporting the hypothesis that the methods for IP address clustering were in fact statistically supported rather than based on an implicit hunch. In fact, Moura verified in a series of experiments that by blacklisting individual hosts only, the spam filter could only block 54% of the spam messages. In another comparison, blocking these individuals plus randomly chosen hosts (not necessarily neighbors) did not improve performance of the system. However, by blocking the individuals plus their neighbors, the detection rate increased to 92%.

The results of the further testing to account for the conflict in values allowed the engineer to address the disconnect in intended versus realized values. In other words, by making explicit that Moura had intended certain values e.g., reliability, effectiveness and fairness, in the Internet bad neighborhood concept these values could then be scrutinized with the future potential deployment in mind. The results of the analysis supported his assumption while at the same time allowed for the researcher to indicate that he had considered the potential disvalue of discrimination but that this threat was also minimized as much as possible. Thus, Moura was able to show that the intended value of Internet security was balanced with the potential threat to fairness, bias or prejudice. These potential minimizations to fairness, bias and prejudice were only made evident through examination with the ethicist.

This value analysis is of course not exhaustive but may be considered the first step in a thorough ethical assessment to facilitate the accompaniment of ethics in ICT research and design processes. What's more, it presents a novel application of the embedded values concept to fundamental research in an ICT institute.

7 Conclusion

The embedded values concept presents the compelling idea that engineers, scientists and designers can create technologies which intentionally enhance cultural and societal values while at the same time minimizing threats to values. Although the embedded values concept (and the resulting design theories that follow) is of great utility, it remains unclear how to utilize this concept in practice. Added to this is the difficulty of utilizing this concept when engaged in fundamental research or experiments rather than in the creation of a commercial product.

The first step in value analysis, as presented here, involves making explicit the intended values of the engineers and/or design team. By making intended values explicit we were able to scrutinize these values i.e. to question whether or not the intended values would be realized in practice. What's more, we also made it clear that although these values may be realized in context the technology also poses a threat to other values like freedom from bias and/or discrimination. Accordingly, the Internet bad neighborhood concept provided an example of the difficulty of embedding values in technology: that the intended values come with value trade- offs and that the technology may be used in a manner that invites other values or dis-values.

Of equal importance is to scrutinize the intended values to ensure that these are in fact the correct values to embed. This too requires in-depth ethical insight as discussed in this paper and was conducted for this research project but was not the focus of this particular paper. Specifically, the goal was to show the difficulty in making values explicit along with the goal of showing the connection between the implicit intended values of engineers and whether these values would be realized in context once the technology was deployed.

References

[1] H. Nissenbaum, "How computer systems embody values," *Computer*, vol. 34, no. 3, pp. 120, 118–119, Mar. 2001.

[2] P. Brey, "Disclosive computer ethics," *SIGCAS Comput. Soc.*, vol. 30, no. 4, pp. 10–16, Dec. 2000.

[3] B. Latour, *Where Are the Missing Masses? The Sociology of a Few Mundane Artifacts*. MIT Press, 1992, pp. 225–258. [Online]. Available: http://www.bruno-latour.fr/articles/article/050.html

[4] A. van Wynsberghe , "Designing care robots for care: Care centered value-sensitive design." *Science and engineering ethics*, 2012. [Online]. Available: http://doc.utwente.nl/84490/

[5] G. C. M. Moura, "Internet Bad Neighborhoods," Ph.D. dissertation, University of Twente, Enschede, The Netherlands, March 2013. [Online]. Available: http://dx.doi.org/10.3990/1.9789036534604

[6] B. Friedman, P. Kahn, and A. Borning, "Chapter 16: Value Sensitive Design and Information Systems," in *Human-Computer Interaction and Management Information Systems: Applications*, 2006, pp. 348–372.

[7] I. Van de Poel and P. Kroes, *Can technology embody values? (to appear)*, P. Kroes and P.-P. Verbee, Eds. Springer Verlag, 2013, vol. Moral agency and technical artefacts.

[8] ——, *Translating values into design requirements (to appear)*, D. E. G. D. Mitchfelder, N. McCarty, Ed. Springer Verlag, 2013, vol. Philosophy and Engineering: Reflections on Practice, Principles and Process.

[9] N. Manders-Huits, "What values in design? the challenge of incorporating moral values into design." *Science and Engineering Ethics*, vol. 17, no. 2, pp. 271–287, 2011.

[10] R. A. Clarke and R. Knake, *Cyber War: The Next Threat to National Security and What to Do About It*. New York, NY, USA: HarperCollins Publishers, 2010.

[11] Cisco Systems, "Cisco IronPort SenderBase Security Network," 05 2012. [Online]. Available: http://www.senderbase.org/home/detail_spam_volume?displayed=last6months&action=&screen=&order=

[12] MAAWG, "Messaging Anti-Abuse Working Group - E-mail Metrics Program: The Network Operator's Perpective – Report # 15," November 2011. [Online]. Available: http://www.maawg.org/sites/maawg/files/news/MAAWG_2011_Q1Q2Q3_Metrics_Report_15.pdf

[13] J. Soma, P. Singer, and J. Hurd, "SPAM Still Pays: The Failure of the CAN-SPAM Act of 2003 and Proposed Legal Solutions," *Harv. J. on Legis.*, vol. 45, pp. 165–619, 2008.

[14] Wikipedia, the free encyclopedia, "CIDR notation," June 2012. [Online]. Available: http://en.wikipedia.org/wiki/CIDR_notation

[15] W. van Wanrooij and A. Pras, "Filtering Spam from Bad Neighborhoods," *International Journal of Network Management*, vol. 20, no. 6, pp. 433–444, November 2010.

[16] S. M. Bellovin, "Security problems in the TCP/IP protocol suite," *SIGCOMM Comput. Commun. Rev.*, vol. 19, no. 2, pp. 32–48, Apr. 1989.

[17] D. A. Wheeler and G. N. Larsen, "Techniques for cyber attack attribution," Institute for Defense Analyses, Alexandria, VA, USA, Tech. Rep., 2003.