International Journal of Technoethics
Volume 13 • Issue 1

The Fairness Impact Assessment:
Conceptualizing Problems of Fairness in Technological Design

Cameron Shelley, University of Waterloo, Canada*

ABSTRACT

As modern life becomes ever more mediated by technology, technology assessment becomes ever more important. Tools that help to anticipate and evaluate social impacts of technological designs are crucial to understanding this relationship. This paper presents an assessment tool called the fairness impact assessment (FIA). For present purposes, fairness refers to conflicts of interest between social groups that result from the configuration of technological designs. In these situations, designs operate in a way such that advantages they provide to one social group impose disadvantages on another. The FIA helps to make clear the nature of these conflicts and possibilities for their resolution. As a broad, qualitative framework, the FIA can be applied more generally than specifically quantitative frameworks currently being explored in the field of machine learning. Though not a formula for solving difficult social issues, the FIA provides a systematic means for the investigation of fairness problems in technology design that are otherwise not always well understood or addressed.

KEYWORDS

Assessment, Design, Equity, Ethics, Fairness, Risk

INTRODUCTION

Walking down a street in an unfamiliar city, a young woman receives an alert on her smartphone. Her *Light Alert* app says that she is in the vicinity of a past assault, as revealed in city police records. The app displays a local map with pins marking sites of past assaults. Feeling cautious, she decides to leave the area.

Light Alert was an app designed by a group of female Indiana University students for Microsoft's Imagine Cup competition in 2010 (Schomer, 2010). Its aim was to help women avoid assaults, a serious and underreported risk especially for college students in unfamiliar cities.

Yet, the app's design makes its operation uncertain. Because the risk of assault in a given place cannot be measured directly, the app predicts it based on proximity to past assaults. Such a prediction is prone to errors: On some occasions, the app will signal an alert when the risk of assault is actually low while, on other occasions, the app will fail to signal an alert when the risk of assault is actually high.

DOI: 10.4018/IJT.291554

*Corresponding Author

Occurrence of these errors indicates a fairness problem, a situation in which the interests of social groups are in conflict. Whereas failing to signal an alert when appropriate goes against the interests of users, signaling an alert when there is no danger goes against the interests of the local community. In the latter case, the community is identified as a place to avoid, to the detriment of its residents. Both groups' interests are legitimate and deserve respect but they are in conflict in the sense that the more one group's interests are realized, the less the other group's are. Here, we face the problem of determining what distribution of interests would be fair.

The purpose of this article is to describe the *Fairness Impact Assessment* (FIA), a framework for characterizing fairness problems and possible resolutions to them. It is important to be clear about what is meant by *fairness*, since this concept has been understood and applied in a variety of senses and contexts (Mulligan, Kroll, Kohli, & Wong, 2019). Here, fairness is taken in the sense of *distributive justice* (Kaufman, 2012), that is, the distribution of burdens and benefits in society. Thus, it does not apply to other problems in which fairness may be applied in other senses, such as the structure of power relationships among social groups (Barabas, Rubinovitz, Doyle, & Dinakar, 2020).

Fairness is an important consideration in technological assessment because, among other things, technology affects the satisfaction of interests amongst social groups (Grunwald, 2009). Where conflicts of interest result, there is a need for them to be settled fairly.

More specifically, technology can give rise to fairness conflicts in at least two ways. First, *endogenous* fairness conflicts are implicit in the design of technology itself, as in the case of *Light Alert*. Second, *exogenous* fairness problems result from the distribution of a technology. For example, if an anti-aging pill were invented but turned out to be very expensive, then only wealthy people may have access to it. Such a distribution could be considered unfair for disproportionately valuing the lives of wealthy people or encouraging the rise of a gerontocracy (Turner, 2003).

It is important to observe, then, that the FIA is not comprehensive in the sense that it does not encompass all the kinds of discriminatory or unjust treatment of social groups that may be mediated through technology. For example, discrimination against social groups may be embedded in social structures and reinforced through technological means (Eubanks, 2018; Hoffmann, 2019). Although the FIA may help to identify such discrimination, addressing unjust social structures is beyond its scope as an assessment of endogenous design concerns.

Concerns about algorithmic fairness have recently become prominent in the field of machine learning (Hutchinson & Mitchell, 2019), where classification systems are used to support socially freighted decisions such as how to allocate healthcare. The FIA can help to provide pertinent fairness assessments in this particular field, such as the depression detection system discussed below. The FIA can also help to relate developments in this field to fairness assessments concerning technological design in general.

The following section describes the FIA, which is a list of four questions intended to help characterize endogenous fairness problems and potential resolutions to them. Addressing these questions provides systematic guidance in recognizing and addressing such problems. The FIA is not an algorithm guaranteed to produce a fair result. It is intended to structure the search for a fair result and to justify its conclusion. Afterwards, examples of the application of the FIA in disparate cases are given, to illustrate its use in a variety of design types.

THE FAIRNESS IMPACT ASSESSMENT

As a framework, the Fairness Impact Assessment consists of the following four questions directed at a given design:

1. What social conflict arises from errors of the design?
2. What constituencies are adversely affected by this conflict? How?
3. What social interests are at stake in this case?
4. How could the conflict be resolved fairly?