Using an *Extended Error Matrix* to promote transdisciplinary collaboration and jointly work towards social justice

Marc Steen, Tjerk Timan, and Steven Vethman; all from TNO

Paper for ESDiT 2022; Track 5: Social Justice and Technology

Keywords: Error matrix, Fairness, Transparency, Transdisciplinary, Collaboration, Boundary object, Boundary spanning

Multiple disciplines are—or would need to be—involved in the design and application of algorithmic decision-making (ADM)[1]. In such a context, it can be challenging for the people involved to find a shared understanding of even basic concepts, like *fairness*. A data scientist, a developer, a lawyer, or a moral philosopher can have rather different understandings of these and other concepts, like *bias*.

We propose to use an *Extended Error Matrix*[2] to promote transdisciplinary collaboration. We speculate that this can function as a *boundary (spanning) object* (Carlile 2002; Carlile 2004; Star 2010) and enable people with diverse backgrounds to better communicate and collaborate, and to jointly work towards social justice.

Below is an example, for the design and application of an ADM system for detecting fraudulent behaviour of citizens, which puts 'orange flags' before the names of people with high risk. The people involved can start with a regular *Error Matrix*—familiar in data science—and, step by step, extend it with new cells, 'outside the box', and discuss topics; see *Figure 1*:

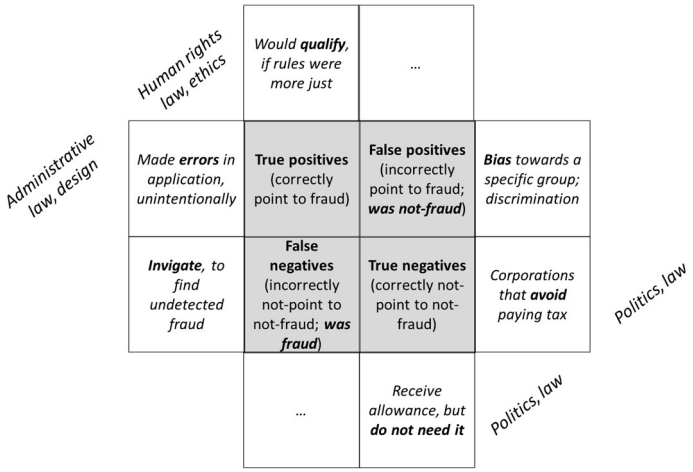

Figure 1: Extended Error Matrix

- **True positives** refer to *correctly pointing to fraudulent behaviour*; we can extend this category as follows:
    - Possibly, this category includes people with good intentions who found the application forms too difficult and made mistakes. Formally, they conducted fraud, but is it fair to punish them for mistakes? This can be a starting point for a dialogue between experts in

[1] This paper results from a transdisciplinary Lorentz workshop on fairness in algorithmic decision-making, in which the authors participated.

[2] Or 'Extended Confusion Matrix', because it is meant to avoid or mitigate *confusion* between disciplines

data science and experts in administrative law (which rules need to be implemented) and human-centred design (how can these forms be improved regarding usability).
    - Possibly, this category includes people who qualify for an allowance, but due to over-stringent rules, they commit 'fraud'. If a family member helps out with groceries, you must notify the authorities—or get punished. Is that fair? This can be a starting point for a dialogue between experts in administrative law, human rights, and moral philosophy.
- **False positives** refer to *incorrectly pointing to fraudulent behaviour*—after investigation, these prove to be *not-fraud*. This happened in the infamous child benefit allowances disaster in the Netherlands. We can extend this category by looking at the following:
    - Possibly, the system is biased towards some type of false positive errors and thereby stigmatizes and discriminates against a specific group of people. Bothering them (again) with (incorrect) accusations, that, after awkward moments and an irritating process of questions, answers, confusion and correction—does not sound very fair.
- **False negatives** refer to *incorrectly not-pointing to not-fraudulent behaviour*—these cases, after investigation, would prove to be fraudulent behaviour.
    - We can extend this category by conducting systematic studies of false negatives, to see whether the system makes some systematic type of error. Such investigations happen very rarely, so that these cases are rarely detected. It would be fair if also not-usual-suspects are scrutinized, whom the system is currently biased towards not-detecting.
- **True negatives** refer to *correctly not-pointing to not-fraudulent behaviour*. Typically, this refers to the majority of people and companies. We can look 'outside the box' and extend as follows:
    - We can look at companies that spend millions on lawyers to find loopholes in tax laws that will save them hundreds of millions. Is it fair to build a fraud detection system that only looks at small fish (citizens who depend on allowances and some of which engage in fraud for thousands of euros) and ignores the big fish (these corporations)? This could be a starting point for a dialogue with policy makers and law makers.
    - We can also look at the rules that enable citizens who *do not really need* this allowance, but nevertheless make them eligible for it. E.g., a reduction on VAT for energy for *all* citizens, regardless of their income or assets. It may be fair to treat citizens equally. But what about equality? This can be a starting point for a dialogue about policies and law.

In the presentation, we plan to share experiences of testing his idea in practice (at TNO). We plan to organize sessions in which we use the *Extended Error Matrix* to promote communication between experts with diverse backgrounds, e.g., data science, law, ethics, and human-centred design.

References

Carlile, P. R. 2004. "Transferring, translating, and transforming: An integrative framework for managing knowledge across boundaries." *Organization Science* 15 (5):555-568. doi: 10.1287/orsc.1040.0094.

Carlile, Paul R. 2002. "A Pragmatic View of Knowledge and Boundaries: Boundary Objects in New Product Development." *Organization Science* 13 (4):442-455. doi: 10.1287/orsc.13.4.442.2953.

Star, Susan Leigh 2010. "This is Not a Boundary Object: Reflections on the Origin of a Concept." *Science, Technology, and Human Values* 35 (5):601-617.