# AI principles in practice: a case by case study

It is not a simple task to put principles into practice, because each project that involves algorithms has a different purpose, with various outcomes relating to new innovations and developments. With Maarten, we discussed the principles of the EU's High Level Expert Group on AI and considered their practical use cases.

## 1. Human agency and oversight

The guidelines specify that humans should be given the knowledge and tools to interact with AI and that humans have certain control over the AI. In case of the algorithms that Maarten developed, a human constantly checks the accuracy, and they can correct certain errors as well: "we monitor how it's performing, if the accuracy is not dropping. We have a human check it, and we use this data to evaluate if it is still doing correctly. If the model is sent to the wrong department, somebody will see it and will send it to the right one. There is a constant feedback loop about its accuracy."

In case of the garbage detection, the aim of the model is also to make this process more efficient, which means that there will be less and less human oversight the more accurate it is: "in the beginning there will be more human oversight, but when it starts performing better we are going to decide to have it less and less." Human oversight is actually something that AI often replaces in practice, since these decisions are often made due to budget restrictions, performance or profit. Ensuring human agency is on the one hand the responsibility of developers in terms of what kind of interface they create, and on the other hand it is the task of all of us to educate ourselves and each other to be more critical of their workings.

## 2. Transparency

The citizen service request algorithm that Maarten developed is a simple verification model, which is easier to explain than other, more complex machine learning systems like deep learning. He wrote an article explaining the algorithm, and the code is available on Github. Even when developers try to be as transparent as possible, there are obvious trade-offs they have to consider: "we can't be completely transparent, because then we would have to share the data which is privacy sensitive," says Maarten.

“The garbage detection model is a newer technique from 2016. Image recognition is more deep learning, so it becomes harder to explain to the public why it is detecting what. This is why I wouldn’t use this technique in predicting fraud that involves personal data, but for scanning garbage it is suitable because you don’t have to explain why it detects what. Currently we do have it recognizing faces, in order to detect them and blur them away.”

Maarten thinks that if it really impacts someone’s life, it is better to use simpler models, because it is necessary to explain why certain decisions are made: “on the other hand, what if you can find more criminals by using deep learning? Is it not ethical to use it?” Decisions about the trade-offs between explainability and efficiency is something that teams working on algorithms must take into consideration, mainly when the algorithms have a high social impact and seriously affect the lives of citizens.

### 3. Technical robustness, data governance

It is not only transparency and privacy decisions where trade-offs have to be made.The guidelines recommend developers to build secure and robust systems that are resilient to attacks, and that personal data is not used unlawfully. Maarten says that “because all the data is really secured, it becomes hard to develop something, because there are a lot of rules of where the data can go, and it becomes harder to make these models more efficient.”

The guidelines also specify that data protocols regarding access to the data should be put in place. At the municipality, civil servants need to sign documents regarding the access to data and its use.

### 4. Diversity, non-discrimination and fairness

Certain risks are involved when we build algorithmic systems in order to have better working, more efficient work processes in many different fields, from healthcare and city management to finance and hiring processes.

How can we ensure fairness and diversity in practice? The guidelines mention that identifiable and discriminating bias should be removed during the collection phase, and that oversight processes need to be in place to analyse and address the system’s purpose, constraints, and requirements in a transparent manner.

Maarten told us about the biases that are present in the citizen service request algorithm. It works more efficiently on Dutch language requests, because the data that the model was trained on mostly contained past citizen requests that were written in Dutch language. So what happens if someone makes a request in English? “I think for bigger classes it might go well, but in most cases it would not be able to detect anything with certainty...so you have the old situation that somebody takes it manually, which does have the bias of if you speak really nice Dutch, your complaint will go to the right department faster and your problem will be solved faster as well,” says Maarten.

This problem could be solved by training the model also on English language data, or by using a translating software during the process. In any case, the request will be taken up manually, so it will not be neglected. The bias that this system holds will not impact the lives of people to a severe degree, but with other systems it can become a serious problem. Austria’s employment agency, which estimates the chances of specific groups on the labour market has demonstrated potential to discriminate against women and disabled, due to the negative weighting of the parameters in the algorithm.

While these cases show that algorithms often contain bias and have the potential to discriminate against certain groups, it is important to highlight the potential of them. Maarten says that we can use algorithms to make the world fairer. “There is an opportunity to train this model to measure if it is doing something that you don’t want it to do and correct it.” In the article ”Discrimination in the age of algorithms”, the authors discuss this issue in a legal context. They say algorithms can be rendered more specific and transparent, than the sort of ambiguity that is present in human decision making: “we can ask exactly which data were made available for training the algorithm (and which were not), as well as the precise objective function that was maximized during the training.”

As we can see, in terms of fairness and discrimination there are both risks and potential benefits related to algorithms, and a task of developers and users is to continue the dialogue, take responsibility for the ethical and social considerations and strive to be more transparent.

### 5. Societal and environmental well-being

The guidelines recommend an environmentally sustainable approach to the use of AI regarding energy consumption. While they acknowledge that AI systems can be used to enhance social skills, they might also contribute to their deterioration. For this reason, there should be careful monitoring. The object detection algorithm at the municipality uses some energy, but as Maarten explained, it is in the interest of the people and the municipality to get something that is not using a lot of energy, because it is cheaper to deploy.

The guidelines point out that the use of AI might impact our social relationships and attachment. While the algorithms that Maarten has been working on do not have a high impact on social relationships, there are more and more applications that might do. Related to this point, we can also argue that television, laptops and smartphones have drastically changed our social relationships already, and AI also has the potential to strengthen social skills.

In this section it is also worth mentioning that these applications often have the potential to replace human labour, therefore it is also crucial to ensure the monitoring of these processes and the reskilling of the workforce if necessary.

## 6. Accountability

Documentation is an important part of accountability, presenting the actions that were taken each step of the way, when needed. At the municipality, “all code is always saved, so it is possible to always look back at what happened exactly. So in the case something happened, you would be able to see what went wrong.”

On the other hand, when we consider accountability and ask who is responsible, it is not completely clear. “Am I really accountable by law if I made this algorithm that is doing something wrong? I would personally feel accountable, because I made it and it is my job so that it works correctly, but I don’t think there are any rules about this. If you look at the bigger level, there is a lot of accountability, different parts of the organization are accountable for how well they handle the citizen service requests, by calculating their customer satisfaction and how well they perform,” says Maarten.

As the guidelines emphasise, humans need to be held accountable and auditing mechanisms need to be implemented. “Blaming an algorithm makes it really easy, according to Maarten. “With the service requests, what you often see is, someone makes a request, something goes wrong and they say the system did it, it was an IT glitch.”

Who do we hold accountable when something goes wrong? Of course it is hard to assign responsibility where the workings of a system are so complex. Do we hold the creator accountable, people that monitor performance, or others that are involved in the process, but not with the technology itself?

According to Brent Mittelstadt, “systems are also often opaque in the sense that no single person will have a full understanding of the system’s design or functionality, or be able to predict its behaviour. Even where problems are recognized, they can rarely be traced back to a single team member or action; responsibility must be assigned across a network of actors that influenced the system’s design, training and configuration.“ Considerations will need to include who assigns these responsibilities and what degree of responsibility we assign to specific roles.

**Final thoughts**

To conclude, we also need to consider in which areas these algorithms are implemented, because there are huge discrepancies. Even though the municipality strives to become more transparent and shares as much as it can regarding the working of these algorithms, businesses often have a contrasting perspective and we would receive different answers to the same questions. The discussion also highlighted that there is a need for people from different disciplines to share their expertise and thoughts, because the field is incredibly diverse. When computer scientists create applications that automate decisions, and they implement them in fields like the media, banking, governments, healthcare and business, it is incredibly important to implement measures that safeguard the rights and well-being of people that are directly or indirectly impacted by these systems.

Author: Petra Biró