START Foundation: Coping with Bias and Fairness when Implementing and Using an AI System

1 Introduction

START-Stiftung gGmbH (“START”) is a German non-profit organization.1 At its core, START provides an education and engagement program for young people with a migration background. By awarding scholarships under the program, the organization aims to promote the educational attainment and increase the social engagement of the young people and thus contribute to an inclusive and fair society (see Figure 1 for START’s logo and exemplary scholarship holders). Over a period of three years, the scholarship holders can therefore attend workshops, events, and seminars with other young people at various locations in Germany to explore their own values and their identity as well as to develop skills that enable them to actively participate in shaping our society.

![START Logo and exemplary scholarship holders](image)

Figure 1. START’s Logo and exemplary scholarship holders

As part of its digital transformation, the organization began to offer part of its education and engagement program online in 2020. Moreover, a digital space called START Campus was developed, where the scholarship holders can exchange ideas and announcements with others in the community.

By the beginning of the year 2022, Farid Bidardel was appointed as the new Chief Executive Officer (CEO). Farid plans to drive digital transformation by creating a Digital Learning Platform (DLP). He intends to expand and optimize the educational and engagement offerings on START Campus and to make it accessible to an even larger target group. In his vision, the DLP will offer Massive Open Online Courses (MOOCs)2, event announcements as well as online mentoring programs besides online networking opportunities.

Given the wide variety of educational and engagement opportunities that START would offer on the DLP in the future, the organization is faced with the need to adopt an approach to manage the plethora of different opportunities the youth are confronted with. Therefore, Farid proposed a project to leverage AI as the use of AI has the potential to assist a virtually unlimited number of users to find useful and relevant information, such as recommendations for courses and mentors, on the DLP (Dong et al., 2020; Gorgoglione, Panniello, & Tuzhilin, 2019; He et al., 2017; S. Zhang, Yao, Sun, & Tay, 2020).

He envisions to develop a digital education and engagement advisor called Skill Compass to prevent information overload, exploit the potential of the DLP and ensure successful education of the young people in the future. In this way, education and engagement can be promoted in an even more targeted manner, and more young people can be reached.

So far, START’s education and engagement program, which is based on on-site events, has been able to support nearly 200 new youth each year. In the future, Farid’s vision is to unlock the potential of more than three million young people with a migration background in Germany through educational and engagement support.

---

1 The data that we report about START in this section can be found on the organization’s website at https://www.start-stiftung.de/, with the exception of information about the Skill Compass. Information is provided in German only.

2 A MOOC is an “online course aimed at large-scale interactive participation and open access via the web”. Information from Conole (2016).
2 Migration Situation in Germany

START had always conducted extensive research on the migration situation in Germany. For a number of years, START’s Development Manager Johanna Bartz has been informing all interested parties in public talks on “German Society – Migration & Integration”. In her last talk, Johanna started to explain:

Migration occurs when a person changes the location of their usual place of residence. International migration occurs when this movement crosses national boundaries.\(^3\) For us, the term migration refers exclusively to international migration; we do not consider migration within a country.

She continued:

An immigrant is someone who moves to a different country. Germany has become one of the most important European destinations for immigrants. Since the early 2000s, the term “people with a migration background” has been commonly used as a collective term for the heterogeneous group of immigrants and their descendants. According to the definition of the German Federal Statistical Office, we consider a person as having a migration background when they or at least one parent does not have German citizenship by law. Therefore, people with a migration background include both those with own migration experience, the so-called first-generation, and people without own migration experience but whose parents migrated to Germany, the so-called second-generation.\(^4\)

Johanna is familiar with the historical development of people with a migration background in Germany (see Figure 2). She has prepared PowerPoint slides that she regularly updates with information from the websites of the German Federal Agency for Civic Education and the German Federal Statistical Office Germany.

![Figure 2. Johanna’s PowerPoint slide about the historical development (Statista, 2022)](image)

Johanna knows that from the early to mid-1990s, there was a high influx of (Spät-)Aussiedler\(^5\), asylum seekers, refugees, and labor migrants. During this time, immigration and integration have become important and highly contested topics in domestic policy discussions. Between 1995 and 2006, the migration surplus decreased. Since 2010, the influx of migrants has been increasing again. In 2015, immigration was at its highest in the history of the Federal Republic of Germany especially due to the large influx of asylum

---

\(^3\) Information from Razum and Spallek (2021).

\(^4\) Information from Razum and Spallek (2021).

\(^5\) Ethnic Germans from Eastern Europe and the former Soviet states.
seekers.6 Over the past few years, the majority of immigrants arrived from European countries, especially from EU member states (German Federal Agency for Civic Education, 2022c).

Johanna said:

In 2020, 21.9 million of the total 81.9 million inhabitants in Germany had a migration background – equivalent to 27 % of the total population. Almost two-third of persons with a migration background came from another European country. The remaining third is composed of persons from various countries outside Europe.7 A German migration expert estimated, that by 2040, about 35 % of the German population will have a migration background.8

She remarked:

The younger the age group, the higher the proportion of people with a migration background. In 2020, among young people aged 15 to 20 years, more than a third had a migration background while the proportion of people with a migration background among senior citizens was only 14 %.9

For START, Johanna researched the challenges for people with a migration background in Germany. She found numerous studies showing German economic assimilation fails due to structural disadvantages in schooling (Dustmann & Glitz, 2011; Gries, Redlin, & Zehra, 2022; Lüdemann & Schwerdt, 2013; Valero, Keller, & Hirschi, 2019). These structural disadvantages are the starting point of START’s work. Hence, she explained in her last talk:

The “PISA shock” at the turn of the millennium revealed that educational inequalities between young people with and without a migration background were more evident in Germany than in any other OECD country. A recent OECD report showed that differences in educational attainment between individuals with and without a migration background remain significant. They reported, that in Germany among persons with a migration background, the proportion without a

6 Refugees from Syria, Iraq, and Afghanistan.
7 Information from German Federal Agency for Civic Education (2022a).
8 Information from Medeiros (2019).
9 Information from German Federal Agency for Civic Education (2022b).
general school-leaving qualification was seven times higher in 2020 than among persons without a migration background.\textsuperscript{10}

In Germany, the Federal Office for Migration and Refugees is responsible for measures for the integration of migrants. Therefore, the Federal Office has developed integration courses and professional language courses. Furthermore, the Federal Office promotes social and professional integration projects that help to strengthen immigrants’ skills and make it possible for them to participate in social life on equal terms and establish a culture of mutual recognition (Federal Office for Migration and Refugees, 2019).

Johanna concluded:

\textit{Despite the existing efforts, access to adequate education for people with a migration background cannot yet be guaranteed. But young people with a migration background are considered a highly ambitious and motivated group for many reasons. They hold great potential to contribute to both the economy and society. START addresses this particular point.}

3 The START Story

3.1 Company Background

\textit{Never doubt that a small group of thoughtful, committed citizens can change the world; indeed, it is the only thing that ever has.}

\begin{flushright}
– Margaret Mead
\end{flushright}

START was initiated in 2002 as a project of the non-profit Hertie Foundation\textsuperscript{11} in Hessen (a state in Germany) with 20 scholarship holders. Following regional expansion, START was transferred to its own foundation in 2007 with the Hertie Foundation remaining the sole shareholder and main sponsor of START Foundation. With a team of 21 employees, 4,000 scholarship holders and alumni from 63 different countries of origin as well as numerous partners from politics, business and civil society, START is nowadays active throughout all of Germany (GHST, 2020; Ministry of Education NRW, 2012)

START advocates for fair education and development opportunities for everyone and thus offers young people with a migration background a chance to improve their education. Their path into and through the German educational landscape was made easier and they were given the opportunity to successfully complete their schooling. In 2018, the foundation sharpened its focus to promote not only the educational attainment, but also increase the social engagement of the young people with a migration background (START, 2022). START’s former CEO Michael Okrob explained:

\textit{The great challenges of our time need a young generation that is as active as possible and aware of its own potential to work for an inclusive society, for sustainable development and for respectful coexistence. Studies show that people with a migration background have so far participated significantly less in politics and civil society than people without. For example, 85 % of people with no migration background voted in the 2017 German elections, but only 65 % of people with a migration background did so. In 2017, almost 40 % of people without a migration background were involved in volunteer work, compared to only 22 % of people with a migration background. These facts are even more alarming because at the same time the idea of democracy is noticeably losing its appeal. Not even one in two young people in Germany under 30 still considers it essential to live in a democracy.}\textsuperscript{12}

He continued:

\textit{If we want to contribute to a better society in the long run, we need to find and encourage those young people who have an attitude of entitlement toward themselves, but also toward others, including the system. We want to find and promote outstanding young people with a migration background in order to achieve improvements for all. Because we see in these young people the}

\textsuperscript{10}Information from Gries et al. (2022).
\textsuperscript{11}With its assets of one billion euros, the non-profit Hertie Foundation constitutes one of the largest private foundations in Germany. Since the establishment, they have dedicated more than 500 million euros to charitable purposes. They support people and projects within the scope of their key issues “Brain Research” and “Strengthening Democracy”.
\textsuperscript{12}Information from The members of the Expert Council (2020); Mounk (2018); START (2022); Heinrich-Böll-Foundation (2018).
ability to change perspectives. Looking at contexts from more than one angle is one of the competencies for the 21st century, in which everything is becoming faster and more complex.13

3.2 The Scholarship

START awards scholarships with a focus on skills transfer, coaching, and financial support.14

The idealistic support focuses on educational and engagement opportunities within the framework of a three-year START curriculum program. In addition to imparting knowledge on the topics of politics, society, culture, media, and STEM15, the non-material support aims to promote individual aptitude-specific potential by strengthening the competencies of communication, creativity, critical thinking, cooperation, and social compass. The cooperation with authorities at the state and local level who support the scholarship holders in the various regions is unique compared to similar programs in Germany—START has so-called regional coordinators in each state, who are trained educators and act as mentors for the scholarship holders. In addition, there are other partners (other foundations, companies, and private individuals) with whom enormous expertise is anchored to help the scholarship holders progress.

Apart from idealistic support, the scholarship holders receive material and financial support: All scholarship recipients receive a notebook at the beginning and annual support of 1,000 euros for individual educational purposes and books.

For the school year 2020/202116, 1,800 young people applied for 190 scholarship places. Together with the scholarship holders from the previous two years, 672 young people were sponsored in that school year. This is related to a budget of four million euros, including 1.5 million euros in targeted third-party funding. Expenditure per scholarship holder amounted to 6,820 euros (see Figure 4) and is mostly attributed to non-material support including more than 400 workshops, events and seminars throughout all of Germany.

![Figure 4. Use of funds per START scholarship holder](image)

Since 2019, the impact of the scholarship has been evaluated. To this end, the agency “Value for Good” (Berlin), which specialized in impact measurement, has been commissioned. Results of a longitudinal study reveal that engagement of young people with a migration background significantly increases during their

---

13 Information from START (2022).
14 The data that we report about START in this section can be found on the organization’s website at https://www.start-stiftung.de/. Information is provided in German only. Applications are open to students with a migration background. They need to attend at least the ninth grade and continue to attend school in Germany for at least three more years. Educational performance, the type of school attended, or the degree sought are not decisive for grant selection. What counts are personality, values, and attitude.
15 Science, Technology, Engineering, and Mathematics.
16 The German school year begins somewhere around late August to early September.
scholarship. For example, all participants experienced an increase in the frequency with which they engage in political discussions in their environment. In addition, participants increase their exchange of ideas with more young people.

A survey among alumni indicated that 97% of alumni graduated from high school, with 34% achieving the best high school diploma grade, 70% started studies at the university, and 49% received a study scholarship.

3.3 Shifting the fellowship program into the digital space

The coronavirus disease (COVID-19) pandemic has impacted virtually all areas of life after it emerged in December 2019. Governments across the world have issued containment and mitigation restrictions to hinder the disease from spreading.17 As a result, START scholarship holders were unable to meet and attend seminars. After an initial feeling of helplessness during the first lockdown in Germany in March 2020, START quickly got down to business and started looking at how it could reach out digitally, bring their young people together despite the lockdown, and give them support and impetus during that difficult time.

In order to stay in contact with the scholarship holders, various talk formats (i.e., STARTlive, STARTxCHANGE, and START2imagin; see Figure 5) were initially developed on Instagram.18 The digital talks were not only open to them, but also to other interested people. This was the first time that the exclusive framework of the scholarship program was loosened. Whether 100 instead of 50 people took part in these digital talks was irrelevant to the budget, but twice as many young people were inspired and encouraged to get involved (START, 2022). START’s Chief Digital Officer (CDO) Gregory Grund remarked:

Such an increase in impact can only be welcomed. Due to the fact that the talk formats on Instagram were so well accepted, the openness for digital formats was strengthened on all sides in the team and among the scholarship holders. Based on this experience, in spring 2020 the idea of a START Campus was born – a digital space for education and engagement for the scholarship holders.

Figure 5. Initial digital education formats

The idea was implemented within few months (START, 2022). Gregory stated:

Since summer 2021, START Campus has been available to the new cohort of scholarship holders for the first time. With their own account, scholarship holders have access to their individual schedule, but most importantly, they can exchange ideas with others in the community. There are thematic and regional group rooms and chat channels. Also, video conferences can be held via

17 Information from OECD (2020).
18 Instagram is a free photo and video sharing app available on iPhone and Android. People can upload photos or videos to our service and share them with their followers or with a select group of friends. Information from Instagram (2022).
the platform, digital whiteboards are available, and documents can be saved. Over time, START Campus will be opened up to other cohorts and alumni. However, these steps are to be implemented cautiously after sufficient experience has been gained.

Figure 6. START Campus

3.4 The Desire to Bring AI and Education Together

On February 1, 2022, Farid Bidardel took the helm at START as the new Chief Executive Officer (CEO). Frank Weise, Chairman of the Hertie Foundation commented:

Farid brings experience working with young people with a migration background, scholarship programs, innovative social entrepreneurship, and fundraising. That are exactly these important skills we require to lead the START program effectively and strongly into the future and to ensure that even more talented young people develop their potential for the benefit of our society.

In his first team meeting, CEO Farid said:

START is now 20 years old, and over 3,500 young people have graduated from the START program, which is something to be proud of. START unfolds its impact approach in the key aspects. But there is still considerable additional scope. Our individual support so far is costly and not accessible to all. With a new perspective from the outside, there is now the opportunity to consider how the impact that START has had for years can be further increased and scaled with technical finesse. So how can we use the possibilities of digitization to increase value to our society?

He has extensively researched the potential of START Campus. He continued:

In the future, our educational and engagement offerings could not only be expanded, but also made accessible to a significantly larger target group. In the future, my vision is to unlock the potential of the more than three million young people with a migration background in Germany. START Campus has so far primarily served as a tool for collaborative learning for our scholarship holders. However, START Campus could be the key impetus for an impact-scaling update of START's operating system. In the long run, everyone who applies to START and even more people might get an account to a digital learning platform.

Farid explained:

Digital learning platform such as LinkedIn Learning, Coursera or Udemy offers a wide range of functionalities – from the administrative control of learning over provision of materials to the delivery of entire courses. They are seen as disruptive innovations that will transform future education. This is because digital learning platforms democratize access to higher education by
making content available online to participants worldwide, regardless of geographic, temporal, or social boundaries.\textsuperscript{19}

Gregory interjected to say:

But finding information and searching for suitable education and engagement opportunities on digital learning platforms can appear unmanageable and overwhelming to users, especially if the platform is constantly growing – which should be our goal, right? This can lead to low usage and high dropout rates from our offerings. We must think one step ahead. We need a tool, that can help users to find relevant items on the digital learning platform, in our case suitable courses, events, peers or even mentors based on their own preferences, skills and visions. A kind of digital education and engagement advisor – a recommender system which provide personalized support to users by learning their previous behaviors and predicting their current needs and preferences.\textsuperscript{20}

Farid concurred and thought out aloud:

You are right and the solution lays in AI technology. With AI we can create advanced insights into what our users need. We can discover comprehensive knowledge in contextual, textual, and visual data. With the help of AI, we can enable users to explore the versatile option space: find individually suitable contents such as top courses people with similar interests have completed or find events related to own recent searches.\textsuperscript{21}

4 Skill Compass for Young People with a Migration Background

CEO Farid and CDO Gregory concretized the project plan and formed a project team. The tool, which should help young people with a migration background to find suitable content on the DLP will be realized in a project starting in early 2023. As part of the pre-planning, they scheduled a two-day conception workshop from 18 August to 19 August 2022 with their team including Farid, Gregory, and Johanna as well as three other employees. The main goal of the conception workshop was to explain goals to the whole team, discuss details of the project planning, and complete the implementation plan for the project while also giving team members time to ask questions. By the end of the workshop, the entire team should have a shared understanding of the functions and opportunities of the proposed AI system.

4.1 First Conception of the Skill Compass

On the first workshop day, Farid started:

Next summer, our digital learning platform START Campus 2.0 will go online. Therefore, I will need your input and your support. In addition to our scholarship program, which will continue as usual, we want to offer our scholarship holders and prospectively all other young people with a migration background in Germany an expanded range of educational and engagement opportunities online. For this, a many thanks to Johanna, who will be leading the development of novel educational and engagement offerings like online courses and digital mentoring programs on the platform.

He continued:

Finding information and searching for suitable educational and engagement opportunities on a digital learning platform can often be unmanageable and overwhelming, especially if the platform has such a broad, diverse, and constantly growing offering as ours will have.\textsuperscript{22} Our goal is to create an adaptive learning environment and avoid information overload for users. Therefore, we need a digital advisor which will be embedded on our digital learning platform – the digital Skill Compass.

He emphasized:

\textsuperscript{19} Information from Faustmann, Kirchner, Lemke, and Monett (2019); Mehta, Chauhan, Gupta, and Jaiswal (2021); Halsbenning and Niemann (2021); Onah, Sinclair, and Boyatt (2014).
\textsuperscript{20} Information from Tarus, Niu, and Khadidja (2017); C.-M. Chen, Lee, and Chen (2005); Onah et al. (2014); Scheibehenne, Greifeneder, and Todd (2010); Khall and Ebner (2014); Zhu et al. (2021); Ataiefard (2017).
\textsuperscript{21} Information from Q. Zhang, Lu, and Jin (2021); B. Chen et al. (2022).
\textsuperscript{22} Information from Zhu et al. (2021); Ataiefard (2017).
We want to design a digital education and engagement advisor which supports young people with a migration background to overcome their challenges and unlock their potential. It should be scalable to many users – prospectively three million young people with a migration background. The Skill Compass is intended to help them find suitable content, access extra-curricular events, and find peers and mentors to enhance exchange and connect with interesting people. To demonstrate our concrete plans, we developed a first mockup.

Figure 7 shows the mockup of the embedded Skill Compass on the DLP. After logging into the DLP, users are confronted with the plethora of different educational and engagement opportunities on the home screen. Placing the call to action button at the top of the home screen, users have the opportunity to activate the Skill Compass and explore their personal education and engagement journey. After clicking on the Skill Compass button, users get a short introduction and are informed about the possibility of personalized recommendations. Thereby, they can check and revise their profile information (demographical data like age, gender, country of origin and place of residence). Users can provide preferences for skills and interests which they can always update in their profile. The recommended items are presented visually and with a textual description to the user. Additionally, each recommendation is accompanied by the strength of support. Further information on the item can be provided upon request. The interface is navigable through the availability of several recommendation lists, e.g., for courses, events, mentors, or peers. By default, a few recommended items are shown, until upon request further recommended items are presented. These items can be filtered and ranked by the user.

Farid continued by providing the team with the technical background, purpose, and potential of recommender systems (RS). Through an embedded RS on the DLP, youth with a migration background should be enabled to explore the option space based on their own preferences. RS became popular in many domains. RS are used to recommend items to purchase in Amazon, they suggest movies or videos to watch on Netflix and YouTube, or they recommend people on LinkedIn to connect with (Gorgoglione et al., 2019; Jugovac & Jannach, 2017). In general, RS help users to find relevant items based on their preferences. Thereby, they support users in refining and exploring the option space according to their preferences (Ricci, Rokach, & Shapira, 2015). Recommendations are generated based on user preferences, features of items, and additional information like past interactions (Isinkaye, Folajimi, & Ojokoh, 2015).
CDO Gregory kept on and explained, that recently, AI has been revolutionizing RS dramatically. The rapid advances of AI are first and foremost attributable to the rise of Machine Learning (ML). ML algorithms are able to effectively capture the non-linear and non-trivial user-item relationships and enable the codification of more complex abstractions as data representations (B. Chen et al., 2022; Mu, 2018). Furthermore, it catches the intricate relationships within the data itself, from abundant accessible data sources such as contextual, textual, and visual information (S. Zhang et al., 2020; Q. Zhang et al., 2021). Recent advances in RS based on AI have gained significant attention by overcoming obstacles of conventional RS – like data sparsity and cold-start (Kumar & Sharma, 2017; Q. Zhang et al., 2021) – and achieving high recommendation quality (S. Zhang et al., 2020).

Gregory said:

We will use AI to recommend courses, events, mentors, and peers. AI allows us to build a recommender system without the need for manual engineering and setting the rules for recommendations ourselves. We can train the system with our own data on educational and engagement offerings. Data of the young people required to generate individual recommendations will collected in real time while using the Skill Compass. The AI-based recommender system will continuously learn from this data and thus promises high-quality recommendations as well as enables cost-effective automatization. We will technically ensure data protection and sovereignty. In the long term, the Skill Compass will be made available open source for subsequent use.

4.2 Creating a Concrete Project Plan

On the second day of the workshop, the goal was to concretize the project plan. The project team agreed that the AI-based Skill Compass should be implemented in three phases:

<table>
<thead>
<tr>
<th>Phase</th>
<th>2023</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Prototyping</td>
<td><img src="image" alt="Graph" /></td>
</tr>
<tr>
<td>1.1 Research on AI methods, User Interface (UI), and Technical Integration (TI)</td>
<td><img src="image" alt="Graph" /></td>
</tr>
<tr>
<td>1.2 Prototyping for AI-based RS and UI</td>
<td><img src="image" alt="Graph" /></td>
</tr>
<tr>
<td>Milestone #1: Proof of Concepts (PoCs) for AI method, UI and TI in DLP</td>
<td><img src="image" alt="Graph" /></td>
</tr>
<tr>
<td>2. Implementation</td>
<td><img src="image" alt="Graph" /></td>
</tr>
<tr>
<td>2.1 Implementation of the RS (backend)</td>
<td><img src="image" alt="Graph" /></td>
</tr>
<tr>
<td>2.2 Iterative development of the interaction concept incl. validation with user testing</td>
<td><img src="image" alt="Graph" /></td>
</tr>
<tr>
<td>Milestone #2: Functional RS (backend) and validated interaction concept</td>
<td><img src="image" alt="Graph" /></td>
</tr>
<tr>
<td>3. Roll-out</td>
<td><img src="image" alt="Graph" /></td>
</tr>
<tr>
<td>3.1 Integration of the RS into the DLP</td>
<td><img src="image" alt="Graph" /></td>
</tr>
<tr>
<td>3.2 Stepwise roll-out for users of the DLP</td>
<td><img src="image" alt="Graph" /></td>
</tr>
<tr>
<td>Milestone #3: Skill compass integrated into the DLP</td>
<td><img src="image" alt="Graph" /></td>
</tr>
</tbody>
</table>

Figure 8. Project plan

The release of the Skill Compass for all users is planned for the beginning of August 2023 as a fully functional version on the DLP. The implementation is scheduled for six months with an expected implementation cost of 250,000 euros. The roles, team members and responsibilities are shown in Table 1.

---

23 ML as a subfield of AI enables machines to find patterns in data sets without explicit programming of rules and to make decisions and predictions based on this analysis. This is made possible by the increasing availability of big data and high computing power. ML applications typically become more accurate the more data they have available – without the need for additional programming. Information from El Naqa and Murphy (2015).

24 Information from Agrebi, Sendi, and Abed (2019).
Concerns about Bias and Fairness

Farid was full of euphoria about the results of the two workshop days. The entire team shared the same vision, and everyone saw the potential and possibilities of the Skill Compass. But Farid was also keenly aware of the debate on the ethical impact and implications of digital technologies (Ashok, Madan, Joha, & Sivarajah, 2022).

Digital technologies – from web-based services to AI solutions – increasingly affect the daily lives of billions of people. There are many hopes but also concerns about their design, development, and deployment (Cath, 2018; Floridi, 2021). Various ethical challenges need to be considered, such as what data should be collected, how long data should be stored, and under what conditions access should be granted (Spiekermann, 2015). Digital ethics experts concern the ethical use of digital technology including privacy and anonymity, responsibility and accountability, transparency and explainability, as well as trust (Floridi, 2021; Kern et al. 2022). The aim of digital ethics is to study and evaluate moral problems related to digital technology usage in order to formulate and support morally good solutions (Floridi and Taddeo 2016; Floridi 2021).

For START, the demanding task is to maximize the value of their digitization process for their scholarship holders and society as a whole, while navigating through ethical issues and legal prohibition (Floridi, 2021). To navigate their digitization process ethically, START must build a framework for digital ethics. Therefore, Mira, a profound expert in digital ethics, started on November 2022 at START.

5.1 Mira’s Onboarding

In her first one-on-one meeting with Farid, he gave her a clear picture of START’s vision and company culture, while also outlining in detail Mira’s tasks at START: Mira should develop a vision and a concrete strategy on digital ethics within START’s digitization process. As one of her first projects, she was asked to accompany the conception and implementation of the Skill Compass. Therefore, Farid started to explain his vision of the embedded Skill Compass on the DLP, told her about the results of the conception workshop and showed the current mockup of the Skill Compass. He noted:

An exceptionally high level of ethics is necessary when interacting with our scholarship holders. This has always been one of START’s great strengths and we do not want to lose it through our digitization process. The Skill Compass is the flagship project in this process. This is why we highly appreciate your role as a guard for digital ethics as part of the Data Science team.

Farid was eager to hear Mira’s initial assessment of the presented AI system. She said:

As a society, we are currently facing multi-layered and complex challenges. The Skill Compass will promise to further leverage the potential of current scholarship holders and makes tailored START offerings accessible to a significantly larger target group – this will help to get many more people active in solving complex societal challenges. That is great! Prima facie, the conception of the Skill Compass also holds great potential to address digital ethics. We can likely ensure full respect for privacy and data protection, adequate data governance mechanisms and legitimized access to data.25

She continued:

---

25 Information from European Comission (2019).
However, I see a potential issue related to bias and unfairness in the AI system. This should be given special focus in this project. Many companies have not yet paid any attention to this aspect. Recently, biased and unfair AI systems have created massive social damage. Amazon, for example, was heavily criticized as they used an AI-based recruiting tool that rated male candidates favorably compared to females. This incident illustrates how an AI system can introduce, reinforce, and indefinitely perpetuate, already rampant unfair and discriminatory practices – in this case gender bias.

Farid commented:

I have heard about this example. Another canonical example for bias in AI systems comes from a tool used by courts in the United States to make pretrial detention and release decisions, right?

Mira affirmed:

Yes, you are right. An AI system named COMPAS was used by the US Justice department to predict recidivism of crime. A study showed that this AI system was biased, and black offenders were seen almost twice as likely as white offenders to be labeled a higher risk but actually did not re-offend. Similar findings have been made in other areas, such as an AI system that judges beauty pageant winners but was biased against darker-skinned contestants. These biased predictions stem from the hidden or neglected biases in AI systems.

Farid took Mira’s first thoughts very seriously. He faltered for a moment. He added:

If we want to implement and use the Skill Compass in next to no time, I need as soon as possible to understand all possible risks regarding bias and fairness. Mira, can you briefly introduce me to the concepts of bias and fairness in AI systems so that I can better assess how we should proceed?

5.2 What are Biased and Unfair AI Systems?

Mira began by explaining the concept of bias:

In general, bias simply refers to a deviation from a standard. We can have, for instance, moral bias in which a judgment deviates from a moral norm. There are many types of bias depending on the type of standard being used.

START needs to focus on the problematic instances of bias in AI systems that may lead to unfair treatment of certain groups of people, implicit discrimination, and perceived unfairness (Köchling & Wehner, 2020). Mira continued:

As for us, we must define bias as an inclination of prejudice towards or against a person, object, or position. Given this definition, I focus on how bias enters AI systems in a first step.

Bias in an AI system can arise at any stage of the AI lifecycle – spanning data collection, development, and deployment of AI systems (Suresh & Guttag, 2021). AI systems require data upon which to be trained. In the cases where the underlying training data contains biases, the AI system trained on them will learn these biases and reflect them into their predictions (Mehrabi et al., 2019). Mira gave an example:

AI systems for image classification, for instance, are often trained on ImageNet, a set of more than 14 million labelled images. But some groups are over-represented, and others are under-represented in this data set – more than a third of ImageNet data comes from the US, home to less than 5% of the world’s population. By contrast, China and India together contribute just 3% of ImageNet data, even though these countries represent more than 36% of the world’s population. This lack of geodiversity led to a proven bias toward Western cultures. An analysis found that an AI system trained on ImageNet labeled a photograph of a traditional US bride...
dressed in white as “bride”, “dress”, “woman” and “wedding”, but a photograph of an Indian bride as “performance art” and “costume”. This type of bias is also known as representation bias. Furthermore, biased AI system decisions might impact interaction between the system and the user. A feedback loop between data, algorithms, and users can perpetuate and even amplify existing sources of bias. She said:

Many AI systems receive feedback when they make predictions. When a search engine serves results, it typically records the links that the user clicks on and how long the user spends on those pages and treats these as implicit signals about which results were found to be most relevant. Such feedback is used to refine the model. But if a user clicked on the first link on a page of search results, is that simply because it was first, or because it was in fact the most relevant for the user? Feedback is tricky to interpret correctly.

She emphasized:

Bias does not only result from data or user interaction. Bias can also result from the way the AI system is modeled or how the system is evaluated. Bias can exist in many shapes and forms. Another common bias type is known as algorithmic bias. Algorithms themselves can display biased behavior due to certain design choices, even if the data itself is not biased. Such as use of certain optimization functions, choices in applying regression models on the data as a whole or considering subgroups, and the general use of statistically biased estimators in algorithms.

The concept of fairness aims to address concerns about bias in AI systems (Bellamy et al., 2018; Feuerriegel, Dolata, & Schwabe, 2020; Saleiro et al., 2019). Mira explained:

One approach to address concerns regarding bias in AI systems is to enforce group fairness, which require that protected groups as indicated by sensitive attributes like race, gender, age or national origin receive similar treatment to others. Adherence to fairness is required from both ethical and legal perspectives. It is forbidden to treat equal social facts unequally or unequal facts equally unless a different approach would be objectively justified.

Mira pointed up:

Unfairness can be intuitively captured through various examples – for instance, when an AI system that approves or denies loan applications has unequal approval rates across genders. Formally defining fairness is comparatively more difficult. Fairness is a complex and multi-faceted concept that depends on context and culture. In the literature one can find more than 20 mathematical definitions of fairness. One widely used definitions is Demographic Parity. Demographic Parity requires that a decision such as accepting or denying a loan application be independent of the protected attribute – 1000 women and 1000 men applicants’ loans are approved. But demographic parity often cripples the utility that one might hope to achieve, especially in the common scenario in which an outcome to be predicated, in our example, whether the loan will default, is correlated with a protected attribute. Demographic Parity would not allow the ideal prediction, namely giving loans exactly to those who will not default.

Mira continued:

The challenge is to select an appropriate fairness definition for a given context to reliably assess the fairness of an AI system. Another widely used fairness definition is, for instance, Equal Opportunity. Equal Opportunity states that each group should get the positive outcome – the preferred decision, such as “getting a loan” – at equal rates, assuming that people in this group qualify for it. This implies that among applicants who are credit-worthy and would have repaid

---

30 Information from Zou and Schiebinger (2018); Suresh and Guttag (2021); Shankar et al. (2017).
31 Information from Mehrabi et al. (2019).
32 Information from Mehrabi et al. (2019); Barocas, Hardt, and Narayanan (2019).
33 Information from Mehrabi et al. (2019).
34 Information from Teodorescu, Morse, Awwad, and Kane (2021); Sen, Dasgupta, and Gupta (2020); Wang et al. (2020); Resnik (2015); Green and Hu (2018).
36 Information from Lee and Floridi (2021) and Hardt, Price, and Srebro (2016).
37 Information from Ruf and Detyniecki (2021).
their loans, both women and men applicants should have similar rate of their loans without specifying any requirement for those that will ultimately default.  

Farid looked concerned. Mira concluded:

*It is of paramount importance to address bias and fairness in AI systems. However, it is not an easy task as bias can occur at any point in the AI lifecycle. Once potential biases are identified and an appropriate fairness definition is selected for the given AI system, several approaches to mitigate bias can be applied, such as building a diverse development team, employing explainable AI methods, or applying bias mitigation algorithms.*  

Unfortunately, there does not exist a “one-size-fits-all” approach. The right answers are constantly evolving as bias and fairness in AI systems is a very active area of research.

6  Shall We Implement and Use the Skill Compass?

After Mira’s explanation about bias and fairness in AI systems, Farid was even more worried. Can the project team mitigate all possible sources of bias, address the concept of fairness in a suitable manner, and ensure an unbiased and fair AI system? May the project team face challenges that they are unable to overcome?  

How should they proceed?

Farid said to Mira:

*We must address all fairness considerations before any rollout. The risk of bias must be reduced as much as possible. We do not want digitization at any cost. The Skill Compass must be in line with our previous impeccable ethical behavior!*

He continued:

*Mira, can you please examine all bias and fairness aspects within the concept of our Skill Compass? I would like to call a meeting even before Christmas exclusively for this topic, where you present your findings to the whole team. Then we will discuss in the team how we should proceed. We will decide together whether and how the Skill Compass should be implemented.*

Farid and Mira agreed that Mira should act like an external consultant to guide the project team through the difficult issue. Together, they determined key questions that Mira should bring to the meeting so that the team could critically discuss the concepts of bias and fairness together and come to an educated decision.

7  Questions

Please develop responses to each question below. You will need to conduct additional research regarding bias and fairness in AI-based recommender systems to adequately respond to each question. Please include references for your research in your responses and state any assumptions that you make.

1) AI-based recommender systems have penetrated every aspect of our lives. AI systems make movie recommendations, suggest products to buy, and who to date. They gain increasing impacts on human and society since a growing number of users use them for information seeking and decision-making.

   a. Describe three benefits to use AI-based recommender systems.
   b. Do you commonly rely on systems that provide product recommendations, such as on Amazon or Netflix? Why or why not?
   c. Would you trust the recommendations of AI-based systems more than those of human experts? Why or why not?

2) Explain how the proposed conception of the Skill Compass align with Farid’s target vision for START. Include the status quo of START in your reasoning.

---

38 Information from Hardt et al. (2016); Lee and Floridi (2021); Cortez (2019).
Further information can be found on http://research.google.com/bigpicture/attacking-discrimination-in-ml/.
39 Information from Bellamy et al. (2018); Mehrabi et al. (2019); Balkir, Kiritchenko, Nejadgholi, and Fraser (2022); Alkhademi, Richardson, Drobina, and Gilbert (2021); Madaio, Stark, Wortman Vaughan, and Wallach (2020); Bird (2022).
40 Information from Mehrabi et al. (2019).
41 Information from Li, Ge, and Zhang (2021).
3) Discuss the relevance of unbiased and fair AI systems for business and society in general and explicitly for START.

4) What aspects of bias and fairness should START consider when developing the Skill Compass?
   a. Who might be treated unfairly? Why?
   b. What types of bias might occur? Name and identify at least three different types of bias.
   c. What are fair recommendations? Draw in your answer also consequences from the problems in the choice of a fairness measure.

5) A number of efforts are being undertaken across governments, nonprofits, and industries, including enforcing regulations to address issues related to bias and fairness in AI systems.\(^\text{42}\) Find and describe two examples.

6) What could START do to mitigate or manage bias in the Skill Compass? Take a stand on the refinement of the project. Among others, you can refer in your answer to the mockup of the embedded Skill Compass on the DLP (see Figure 7) as well as to the project team’s roles and responsibilities (see Table 1).

7) Should START embed the Skill Compass on the DLP?

8) Can the Skill Compass also be a model for the use of education for other contexts?

\(^{42}\) Information from Srinivasan and Chander (2021).
References


START Foundation: Coping with Bias and Fairness when Implementing and Using an AI System


