



**ALGORITMI
+ INCLUSIVI**

Artificial Intelligence and LGBTQIA+ inclusion

Report abstract

April 2026



LGBTQIA+ LEADERS FOR CHANGE



| | |
|---|--------------------|
| Foreword | 01 |
| The Report and this abstract | 02 |
| Algorithmic Biases | 03 |
| The uses of Artificial Intelligence and discrimination | 05 |
| Artificial Intelligence and regulatory frameworks | 06 |
| For a solution-oriented ethical approach to concrete problems | 07 |

Foreword

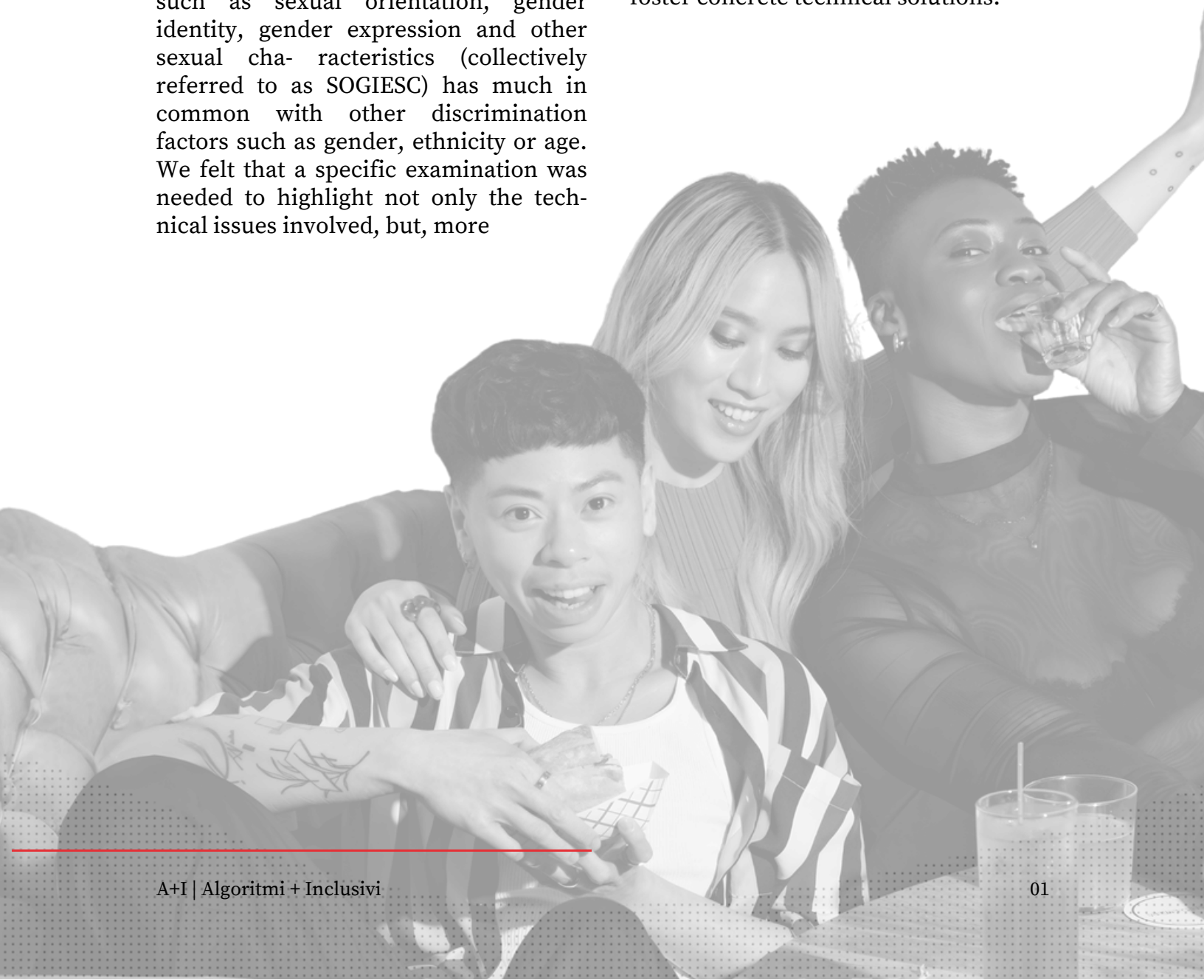
In 2021 [EDGE LGBTI+ Leaders for Change](#) began to explore the potential impact of artificial intelligence on discrimination and inclusion among LGBTQIA+ people and other vulnerable groups.

Reports of discrimination linked to the use of artificial intelligence (AI) tools were beginning to circulate concerning, for instance, discrimination against non-white people in facial recognition systems, against women in automated rankings and against minorities in criminal recidivism risk assessments.

The relationship between AI and factors such as sexual orientation, gender identity, gender expression and other sexual characteristics (collectively referred to as SOGIESC) has much in common with other discrimination factors such as gender, ethnicity or age. We felt that a specific examination was needed to highlight not only the technical issues involved, but, more

importantly, their impact on people, underscoring the need for the diversity - that we as people embody - to be meaningfully considered in the design and deployment of AI systems, with a particular focus on SOGIESC - related factors.

We therefore decided to launch a study and research project aimed at disseminating information and raising awareness in order to understand whether and how researchers, businesses and public institutions are addressing the issue, focussing on aspects related to sexual orientation and gender identity, so as to assess whether it is possible to foster concrete technical solutions.



The Report and this abstract

The overall Project is called ***“A+I Algoritmi + Inclusivi”*** (***“A+I More Inclusive Algorithms”***), which has several strands of activity and within which we have created a multidisciplinary working group.

Among other activities, we conducted interviews and dialogues with experts, researchers, practitioners and key players in the development of AI as well as a series of dialogues at conferences and business meetings.

This has resulted in a **Report** (of which this document is a concise preview) that builds on those interviews and expanding its scope with an overview of the core concepts, seeking to provide a more comprehensive presentation of the issues with the greatest impact on the relationship between AI and discrimination.

What emerged from both the interviews and our review of the literature is the extent to which discrimination against LGBTQIA+ people and their inclusion are marginalised within both research into AI and broader reflection on the subject.

This is confirmed by the later stages of the research. For example, a search on the [Scopus](#) database using the keywords ***“artificial intelligence”***, ***“gender”***, ***“sexuality”***, ***“sexual”*** and ***“fairness”*** yielded 230 results.

Only nine contain the expressions ***“sexual preference”***, ***“sexual orientation”***, ***“sexual minorities”***, ***“gender identity”***, ***“queer”*** in the abstract.

Of these nine, only three include ***“gender identity”***, one ***“queer”*** and none ***“LGBT”***.

At least from a European perspective, one might assume that the limited number of studies on the relationship between AI and discrimination or inclusion of LGBTQIA+ people stems from a very limited ability to collect and use data on people's sexual orientation, gender identity, gender expression and other sexual characteristics.

This explanation is unconvincing, not only because artificial intelligence systems are capable of inferring protected characteristics from other data points or by correlating data, but also because there are many issues that have a direct and distinctive impact specifically related to SOGIESC (Sexual Orientation, Gender Identity, Gender Expression and Sexual Characteristics) factors.

Particularly striking are projects that explicitly seek to identify people's sexual orientation, for example through [facial recognition](#).

In this context, the accuracy of the results is largely beside the point: the harm arises from the very use of such systems.

More broadly, both the potentially harmful and the potentially innovative dimensions of AI are defined by the impact of these systems.

Algorithmic Biases

Any meaningful reflection must, of course, begin with the issue of so-called **algorithmic bias**, a term with multiple nuances that can be understood broadly to encompass not only **algorithmic biases** in the strict sense, but more generally the ways in which artificial intelligence systems reproduce and amplify stereotypes and generate discriminatory output.

In January 2023, [a Wired campaign](#) showed how an image-generating AI responded to prompts asking people to imagine “**two lovers**”, “**manager**”, “**athlete**”, “**parents**” and so on. The result was largely stereotypical. In the context we are dealing with, couples and families were always depicted as heterosexual. Wired's video made the issue strikingly clear, exposing a problem with representation and the erasure of people excluded by prevailing stereotypes. Two years after Wired's campaign, image-generating systems are producing more varied outputs, but they remain difficult to fully control through prompting.

For other examples, simply try and prompt unfiltered generative language models that will reveal their potential to reproduce stereotypes, fake news, offensive expressions, etc.

Debora Nozza, Federico Bianchi and Dirk Hovy, from the Department of Computer Science at Bocconi University, studied the [BERT language model](#), developed by Google between 2018 and 2019.

BERT is (was) a model used by Google to understand the meaning of a word based on context. The researchers prompted BERT to complete a series of sentences written in different languages, measuring

the likelihood that the model would complete them with offensive language, both in general and specifically targeting women and the LGBTQIA+ community.

BERT produced offensive results for all identities, but a particular gender **bias** emerged, as 4% of male subject sentences, in contrast to 9% of female subject sentences, were completed using sexual references. In contrast, when the sentence involved an LGBTQIA+ subject, completions that were sexually explicit or offensive averaged 13%, with peaks of as much as 87%.

Numerous other studies have analysed and confirmed the association between terms related to discrimination factors and offensive or otherwise discriminatory terms.

AI models, in fact, absorb the stereotypes present in the data with which they are created and trained. Additionally, they can absorb the **biases** of the people involved in their design, training and testing, without any preliminary consideration of these influences.

The problem, however, is not limited to the mere reproduction of **biases** in algorithmic outputs, as the models themselves can amplify the **biases** present in the data. The so-called “black box” effect, i.e. the limited ability to understand the operation of AI models and in particular the relationship between model input and output, is relevant in this perspective.

Discussions held during the preparation of the Report suggest that the AI developer community tends to perceive **bias** primarily as a technical problem related to the accuracy of results.

Addressing this problem remains delicate, often involving choices that are inevitably

shaped by the different world views, knowledge, sensitivities and skills of those involved in the process.

Numerous attempts have been made to find technical solutions to mitigate the impact of **bias** in the data without appreciably decreasing the accuracy of the models. **A relevant fact is, however, the current limited [level of ethical skills](#) within the teams of AI developers.**

It is worth mentioning here that the amplified surfacing of **biases** conveyed through the use of language also allows an analysis of language itself, highlighting the **biases** in the common use of verbal and non-verbal communication tools. In other words, studying how **bias** is amplified allows us to backtrack and analyse the cultural features of the language with which models were trained.

When considering the impact of AI on discrimination and inclusion, it is important to remember that there is no single form of artificial intelligence: rather, there are many models and many systems, developed with different approaches, functions and uses.

A taxonomy developed by the [Joint Research Center \(JRC\)](#) of the European Commission in 2025 identifies the main sub-domains of AI development, offering a comprehensive view of the field. These include knowledge representation, automated reasoning, planning, search, optimisation, machine learning, natural language processing, computer vision, sound processing, multi-agent systems, robotics and automation, self-driving vehicles (as well as other areas less relevant here), to which broader content generation should be added. One must then consider also how these domains combine.

The reflections outlined here with reference to generative artificial intelligence, widely encountered and used since late 2022, should be extended to all algorithmic systems, from image recognition to content moderation and beyond. Systems used in the digital sphere, in particular those involved in content recommendation and moderation, are especially striking in this regard.

A concrete example may be useful. Some years ago, a major social media platform temporarily blocked the publication of kisses between men. This occurred because, during the development and training of the AI system used for automated content moderation – a production phase carried out in a cultural context different from our own – images of two men kissing had been technically **labelled** as inappropriate. The restriction was eventually lifted, but only after numerous users' complaints.

A similar case, this time reportedly the result of a deliberate policy choice, was attributed to TikTok by the developers of the dating apps Surge and Zoe, who [publicly objected](#) to the permanent suspension of their account after the publication of a photo depicting a kiss between two women that allegedly violated the policy on “**intimate kissing**”. The developers pointed out that similar content depicting heterosexual kissing were not removed, highlighting a double standard in content moderation that discriminated against same-sex relationships.

These cases show that AI-related **biases** exist along a continuum: from implicit **biases** embedded in data sources (such as systemic **biases**) to **biases** arising from data structure or collection methods (such as measurement or sampling **biases**), and finally to **biases** directly introduced by

people involved in model development, both in relation to the data entered during the various stages of the process and in relation to the more specific algorithmic parameters and objectives.

The uses of Artificial Intelligence and discrimination

The enormous potential of algorithmic systems means that they have or can have extremely wide-ranging impacts across geographical and social contexts, as well as a pervasive influence.

These systems may, to varying degrees, embed stereotypes and discriminatory practices, and the existence of such **biases** can weigh differently in decisions about whether to deploy a given algorithmic system. **However, there are cases where AI is used in a way that deliberately seeks to discriminate.** As noted, the digital environment is one in which automated decision-making is already widespread and can specifically impact LGBTQIA+ people.

A [recent study](#) in China analysed how social media content posted by gay men is subject to so-called **shadow banning**, a form of moderation that reduces the visibility of content deemed inappropriate on these platforms.

Moderation can take several forms.

For example, it may occur by (i) substituting homophones (words that are pronounced the same way but have different meanings and often different spellings) in the

search field, thus effectively preventing the search for keywords considered inappropriate, (ii) displaying messages suggesting that, under the law, no results can be provided for a given search, or (iii) replacing certain words with synonyms (for instance, substituting the word “**gay**” with “**comrade**”).

In addition, (iv) the proposed results are often characterised by a high level of so-called noise, requiring users to scroll through many pages before finding what they were looking for.

Moderation (v) also operates at the level of hashtags, where unwanted ones are not removed but cannot be used to search for related content. Content dissemination is also manipulated: on video platforms, content about gay men is prevented from exceeding 100,000 views.

There are even more sophisticated and no less powerful approaches designed to induce **bias** and negative effects on people. This can be done by attacking an algorithmic system using commands designed to alter its behaviour, or by supplying inaccurate data during the development and training phase of a model. Such practices are generally aimed at inducing specific, undesirable outcomes (so-called **poisoning**). One example involves the large-scale dissemination of online data deliberately intended to be harvested by the automated systems that populate the datasets used to train **LLMs** (Large Language Models).

These are not hypothetical risks but well-known vulnerabilities that developers of algorithmic systems are constantly trying to remedy, but which are widely exploited by malicious actors.

The same database-distorting effect may also result from the actions of many US government agencies and entities (including many entities in charge of health and social research) which, as a result of certain executive orders issued by the Trump administration in early 2025, are removing documents from the web, as well as deleting information or individual words from the available documents, such as the term “*gender identity*”, the use of which was essentially banned by [Executive Order 14168](#).

The scientific community, companies and lately policy makers are seeking to address the problems outlined above with methods and tools grouped under **the acronym FAccT (Fairness, Accountability, Transparency)**, a strand of work that is more than 15 years old. However, there are no simple, fully neutral or universally valid solutions capable of addressing all forms of discrimination across every context. Moreover, the measures adopted are never final, as they must be continuously revisited and updated as models evolve.

Some issues are particularly complex, such as (i) how to sample so-called “*unobserved*” characteristics like sexual orientation, (ii) whether sensitive data, such as SOGIESC-related information, should be processed to improve the accuracy of AI systems, and (iii) which methods of identifying and correcting biased outcomes are acceptable. These are issues that demand reflection and decisive action, otherwise we'll be overtaken by events.

Artificial Intelligence and regulatory frameworks

Finally, the role of legislators cannot be overlooked. Implementing principles of equality, equal treatment and non-discrimination in the context of artificial intelligence poses new challenges, given the potential conflict between those principles and others.

The **European Union** is particularly active in this area. It has established the **General Data Protection Regulation (GDPR)**, a global benchmark for personal data protection, and developed additional regulations with significant implications for safeguarding individuals in their interactions with AI tools such as, in particular, the [Digital Services Act \(DSA\)](#) of 2022 and the [AI Act](#) of 2024.

The [GDPR](#), introduced in 2016 as Europe's data protection and privacy standard, has demonstrated its effectiveness, albeit with certain limitations, in addressing numerous cases of automated processing that can violate individual rights. Yet the scale of system development makes the limited means available to the authorities apparent. The DSA is highly relevant to the regulation of digital services and the moderation of online content.

The AI Act, on the other hand, is crucial for enforcing the foundational principles that govern the use of AI in highly sensitive domains where the risks of discrimination are particularly significant. We most notably see that in employment and access to essential services, but also in areas such as biometric recognition, the administration of justice, criminal prosecution and certain

forms of social scoring. However, in implementing both regulations there is a risk that the protection of vulnerable groups covered by anti-discrimination laws may be overlooked amid the sheer volume of issues at hand.

From a legal perspective, it is also important to underline the risk that discriminatory practices amounting to direct discrimination may be reclassified as indirect discrimination, and thus subjected to a different legal standard, simply because they are carried out by algorithmic systems. Such an outcome is clearly unacceptable, as it weakens the protection of individual rights. Constant awareness-raising and advocacy, including by civil society organisations, and verification of implementation are essential.

EDGE's "**A+I Algoritmi+Inclusivi**" (**More Inclusive Algorithms**) Project, of which this abstract and the Report are an integral part, embodies this perspective and purpose.

Furthermore, there is significant scope for AI applications that fall outside high-risk categories, for which it is essential to develop and disseminate tools that promote inclusion and are grounded in a human-centred vision of AI. This vision is not only ethically sound but also commercially prudent, as people must be able to trust and rely on these tools.

Strategies and policy tools can be identified to reduce the risk of discrimination in relation to these themes, including participatory design approaches and efforts to raise political awareness, supported by close collaboration with expert committees and corporate ethics boards.

Clearly, neither this abstract nor the Report can be exhaustive. The evolution of AI-

based tools continuously raises new questions, from biometric technologies that may have particularly significant impacts on trans and non-binary people, to the relationships people "**form**" with chatbots, which can expose vulnerable people to risks ranging from outing to mental health issues and so on, to the microtargeting of algorithmic advertising, which can prove exclusionary or highly intrusive, to the use of AI in public policy, such as welfare systems, which risks reproducing decades of institutional discrimination embedded in historical data.

For a solution-oriented ethical approach to concrete problems

EDGE, an association for civic activism and rights that has been active for over twelve years in Italy and within a European network of like-minded organisations, with a mission to promote LGBTQIA+ inclusion in the workplace, the professions and business, has consistently taken a proactive approach and pursued pragmatic solutions.

The question, therefore, is what can be done. The scope for intervention is significantly constrained. Interacting with AI development requires considerable resources in terms of skills, people, data, computational power and so on. We have sought to outline a concise set of priority areas for intervention and possible courses of action.

1) First and foremost, it is essential that both users of AI tools and those affected by them are aware – among the

many new insights related to AI – of the potential risks of discrimination inherent in the use of AI tools and systems.

In particular, we believe that the LGBTQIA+ community should take the lead in fostering this awareness internally, developing practical tools for protection and proactive advocacy, and in sharing them both with individual potential targets of discrimination and with its allies.

At the same time, there is a clear need for widespread awareness of *bias* and discrimination across all levels of management, from technical to administrative, both within companies that develop AI technologies and in those that deploy them in their products, services or operations. This is necessary not only to ensure regulatory compliance, but also to promote ethically sound and professionally responsible use of the tools, with particular attention to areas such as human resources and social media content production.

2) It is also essential to support academic and business initiatives that study discrimination in the field of AI and to foster new ones. There is also a need for more research into the algorithmic discrimination of LGBTQIA+ people and the delicate trade-offs that need to be resolved in developing tools to address the problem across the technical spectrum.

3) Strong collaboration should also be sought with civil society organisations working on digital rights and with the so-called **trusted flaggers** accredited under the Digital Services Act.

4) Another key priority is to oversee the implementation of the AI Act by participating in the relevant processes. Particularly in Italy, it will be necessary to establish channels for interaction with

AGID (the Agency for Digital Italy) and ACN (the Agency for National Cybersecurity) identified as the Italian AI authorities but historically not involved in anti-discrimination issues.

Equally important is participation in the development of the harmonised standards required by the AI Act.

It will also be necessary to monitor the proposed amendments to the GDPR and the AI Act that are currently under discussion.

5) Associations, consultants and people involved in workplace inclusion also have a crucial role to play, particularly with regard to the implementation of AI tools for human resources management.

The first [toolkit](#) created for this purpose already exist. **Combining the expertise and perspectives of the people working on inclusion with the people developing and implementing AI systems, particularly in the world of work, is crucial.**

In particular, more structured companies that are developing processes for the deployment and evaluation of AI tools should involve ERG/BRGs (Employees/Business Resource Groups) in these processes, particularly in impact assessments, including the Fundamental Rights Impact Assessment.

6) Finally, it is important to remember that, given the increasing adoption of AI tools, it is both possible and necessary to experiment with the positive uses of AI to foster LGBTQIA+ inclusion.

This is a constantly evolving plan of action.

The “A+I Algoritmi+Inclusivi” Report by EDGE

This abstract constitutes a summary and preview of the “**A+I Algoritmi+Inclusivi**” (**More Inclusive Algorithms**) Report, which will be published in 2026.

It introduces the concepts needed to understand artificial intelligence and presents a roadmap of key issues related to the risks of algorithmic discrimination.

The Report examines the principal risks of discrimination affecting minorities, particularly the LGBTQIA+ community, within AI systems. We use concrete examples and analyse the main problems, starting with the black box effect, and considering the major areas of work, foremost among them fairness. It also surveys techniques for improving AI explainability and mitigating **bias** through the selection and manipulation of **datasets**. Finally, the issue of discrimination is explored in some specific and high-impact applications, including natural language processing and automated image recognition.

The Report goes on to examine the business landscape, addressing the levels of awareness around the risks of algorithmic discrimination and efforts made to reduce it. It underscores the need to foster a widespread culture of understanding around **bias** and discrimination issues at various levels, from the technical to the administrative, both in companies involved in producing AI technologies and in those using them. Particular attention is paid to the risks of data processing and profiling, as well as to specific challenges in the areas of recruitment and social media content production.

The Report also attempts to approach discrimination from a humanities and social sciences perspective, distinguishing between **weak** and **strong** AI, and between very long-term concerns and a more pragmatic focus on the immediate, current and concrete risks of discrimination and disempowerment, including those related to the datafication of sexual identity.

It explores the key features of human interaction with machine learning algorithms and then moves on to a reflection on the importance of understanding how artificial intelligence works and the risks it poses across society at large, from civic life to politics.

The discussion then returns to the concept of **bias**, the trade-off between efficiency and explainability, proposing an integrated approach that interprets machine learning as machine socialisation, and **bias** as a form of **habitus**, moving beyond a purely quantitative point of view. The Report then examines the main European regulations addressing discrimination and AI, focusing on the GDPR, DSA and AI Act.

Finally, it outlines the desirable characteristics of AI systems designed to minimise the risk of discrimination within the framework of the European AI Act, identifying key areas for action to safeguard the rights and inclusion of LGBTQIA+ people.

The digital version of this abstract, including links to the various sources cited, together with the full report, is available on the EDGE website. Frame the QR code.



EDGE

EDGE's "**A+I Algoritmi+Inclusivi**" project was coordinated by a team consisting of **Mario Di Carlo, Alessandra Galli, Damiano Terziotti** and **Luca Trevisan**, who left us all too soon in 2024 and to whom EDGE dedicated the AI4Good award together with Luca's family and the Department of Computer Science at Bocconi University.

Contributors to the project were **Arsenico – La nuova comunicazione** for creativity and communication coordination, **The Fab Lab** for interviews, **Artdisk | Design & Comunicazione Fluida** for the presentation and layout of the report.

We would like to thank, for taking part in the interviews, **Massimo Airoidi, Luca Altieri, Luca Belli, Alessandro Bonaita, Silvia Franzecco, Luisella Giani, Dirk Hovy, Gianclaudio Malgeri, Pietro Monari, Debora Nozza, Dino Pedreschi, Stefano Quintarelli, Omer Reingold, Bruno Ronsivalle, Luca Trevisan** and the many friends and acquaintances we had the opportunity to talk to, with a special mention for the guests of our Artificial Cafés: **Brando Benifei, Federico Cabitza, Mia Caielli, Alessandro Castelnovo, Alessio De Luca, Davide Locatelli, Tommaso Mauro, Rosa Meo, Chiara Natali, Guido Noto La Diega, Ilaria Penco, Francesco Rizzi, Federico Sartore, Roberta Savella, Vincenzo Tiani, Roberto Trasarti.**

Special thanks to **Silvano Bertossa** for his invaluable support in closing the report and his enormous patience.

The image used, licensed under Creative Commons (CC BY-NC-ND 4.0), is part of the project created by VICE [«The Gender Spectrum Collection»](#), a photo library of images depicting transgender and non-binary people in real-life contexts.