European Commission

HUMAINT

Understanding the impact of Artificial Intelligence on human behaviour

CAS Centre for Advanced Studies

Joint Research Centre

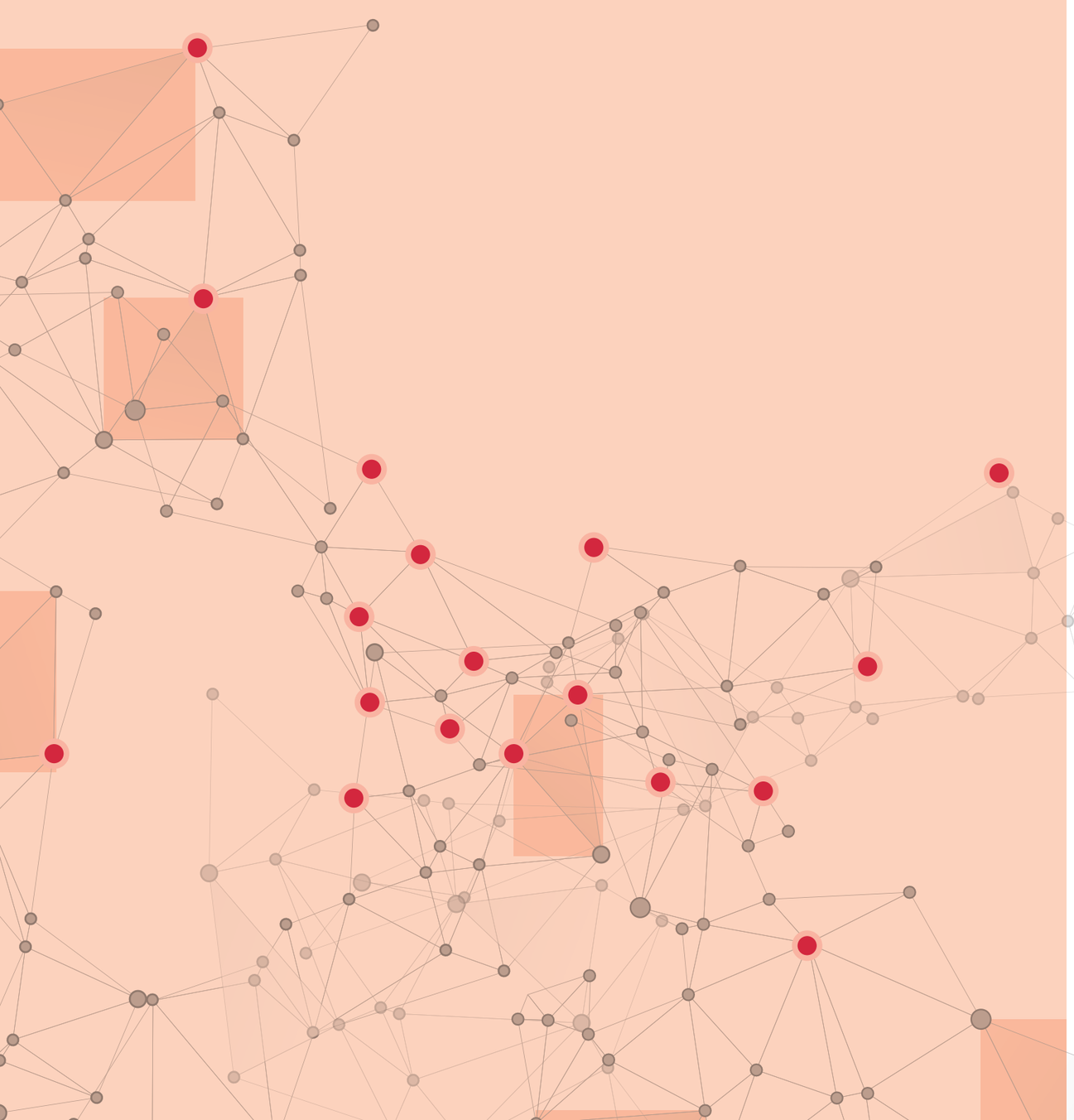


TABLE OF CONTENTS

Centre for Advanced Studies 5

What is the HUMAINT project? 7

What causes algorithmic bias in criminal decision making? 8

The occupational impact of AI 19

Social robots and human development 23

AI for music creation and listening 31

AI in medicine and healthcare 35

Diversity in AI 38

Emergent ethical considerations and community building . 40

Conclusions 42

Who are HUMAINT? 44

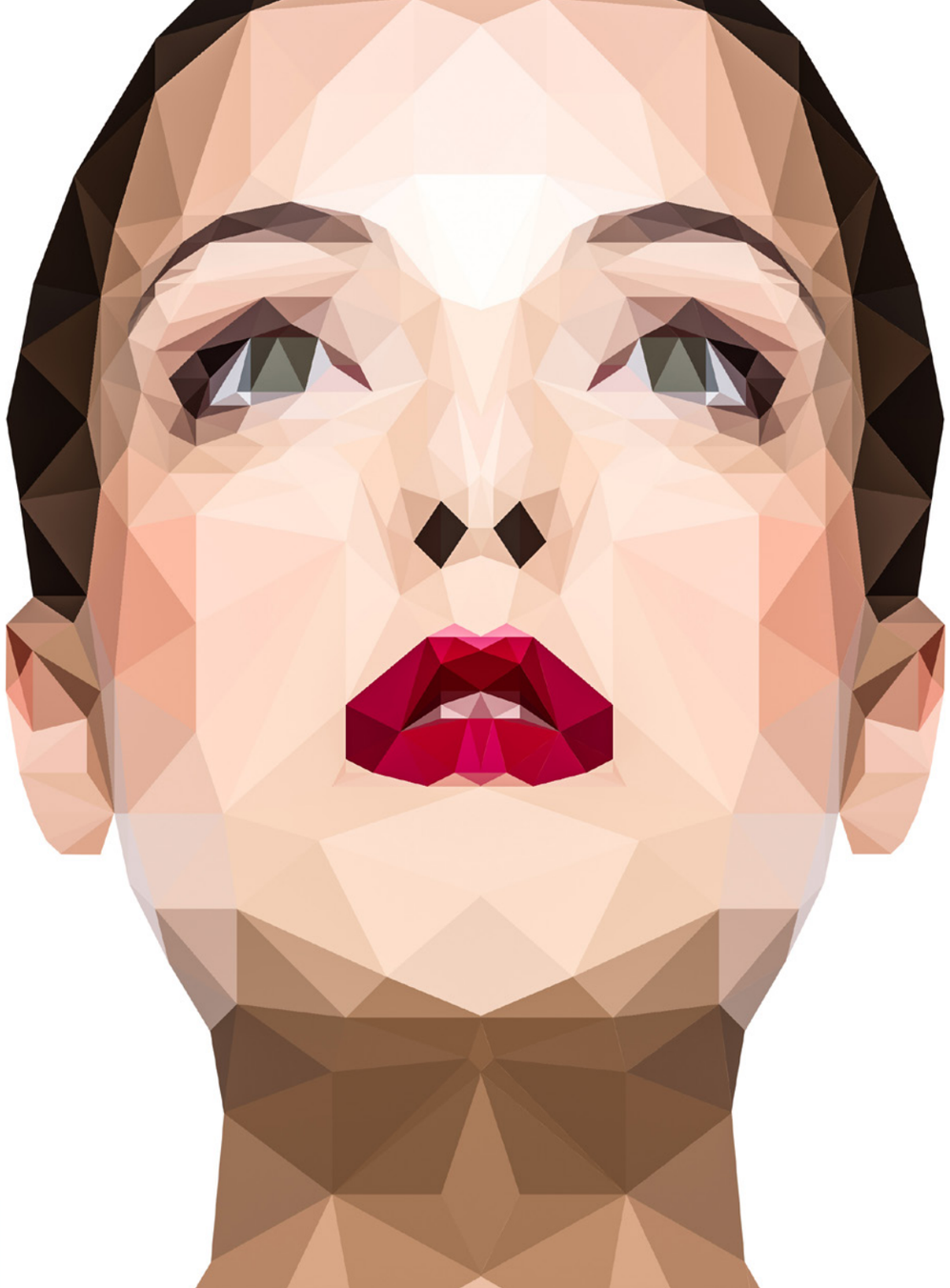
References 44

Centre for Advanced Studies

The Joint Research Centre (JRC) of the European Commission carries out research in order to provide independent scientific advice and support to EU policy. The JRC's Centre for Advanced Studies (CAS) was created in 2016 to give the JRC a leading edge on societal issues that may become relevant for EU policy making and to assist societies as a whole.

By creating the conditions necessary for innovative and interdisciplinary research, as well as offering a creative and generative space in which ideas and knowledge in emerging thematic fields across different scientific and technological disciplines can thrive and flourish, CAS has become an incubator for cutting edge research, formal inquiry, stimulating ideas and activities. It provides the JRC with new insights, data projections and solutions for the increasingly complex medium and long-term challenges facing the EU. So far topics addressed included artificial intelligence, demography, big data and digital transformation.

The CAS project, HUMAINT (Human Behaviour and Machine Intelligence) began in 2017 and will be completed in 2020. This brochure provides an overview of the project and its main achievements.



What is the HUMAINT project?

Artificial intelligence (AI) systems, when applied in practical applications, have an impact on human behaviour. On the one hand, AI provides cognitive assistance to humans, such as helping us to interpret data more efficiently and discover hidden knowledge in large data resources. On the other hand, these AI systems may also affect human decision making and cognitive tasks.

The goal of the of the HUMAINT project[1], is to advance the scientific understanding of the impact that AI systems have on human behaviour. Our research has three main characteristics. First, it is interdisciplinary: we combine different methodologies such as behavioural studies, machine learning practices and statistical methods; second, it is reproducible, as we generate open publications, datasets, code and research protocols; third, it is collaborative, as our work is carried out in the context of networks and partnerships with different research institutions.

In order to have a comprehensive understanding of the impact of AI on human behaviour, our research touches upon different sectors of society where AI may have a particularly large social impact, such as machine learning algorithms in decision making in the criminal justice system; in our work lives; in our social interactions; in music creation and listening; in medicine and healthcare; and in diversity. We then establish commonalities between these scenarios to arrive at scientific and policy-relevant conclusions. Through our research, we also promote value-centered and ethical approaches for the development and application of AI systems.

In the following sections, we present these different scenarios, the methodologies we use, and the main challenges that we are currently addressing.

1 Human behaviour and machine intelligence https://ec.europa.eu/jrc/communities/community/humaint

What causes algorithmic bias in criminal decision making?

In this scenario, we focus on the causes of discrimination and the potential sources of biases that may occur in different parts of the decision-making process in the criminal justice system.

Indeed, it is not new that algorithms, just like humans, are biased in their decisions [25]. This is because algorithms derive their decision rules from data, and we obtain data from past human decisions that are shaped by institutions and societal rules.

So, algorithms can inherit biases from any stage of this decision process through data, as demonstrated in Figure 1. In addition, developers who are biased themselves can introduce biases into the algorithm during development. Finally, algorithms can perpetuate these inherited biases by repeating them in new deci-

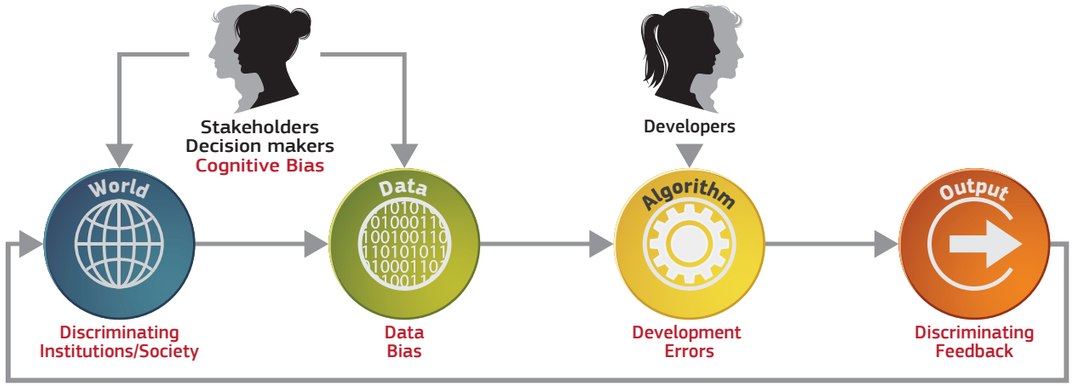


Figure 1: Bias in algorithmic decision **making**

sions at a large scale and creating new biased data. We illustrate these sources of bias in [25, 32].

Our main case study is assessing the recidivism risk of defendants in Catalonia. In this context, recidivism is defined as the act of a person committing a crime after they have been convicted of an earlier crime. We compare how biased humans are compared to machines in predicting the likelihood of reoffending. We use information and data from legal judgements, which were available via open data sources [26]. To assess how effective AI is at predicting the risk of recidivism, and whether it is fair (i.e. the machines do not discriminate against race or sex), we compare the risk assessment tool Structured Assessment of Violence Risk in Youth (SAVRY) to several machine learning models. We wanted to see if machine learning models are better in predicting recidivism and if they exhibit any discrimination, in comparison to SAVRY.

In addition, we use interpretable machine learning to trace back discrimination by assessing which features were important when classifying a defendant as "high risk". Finally,

> "We compare how biased humans are compared to machines in predicting the likelihood of reoffending."

we explore in more detail two possible causes of algorithmic bias in the data with respect to the recidivism of the two protected groups, which are foreigners (more specifically non-Spanish nationals) and females. In particular, we look at (1) the differences in the prevalence of re-offending between these protected groups and (2) the use of the group features (sex, race) or features correlated with them in the training of the algorithm. Our analysis shows that both (1) and (2) can lead to discrimination. We observe that using methods to mitigate the influence of either cause do not guarantee fair outcomes [19].

Machine learning systems to assist judges in predicting recidivism

When judges decide whether to detain or release defendants awaiting trial, they must consider the risk of the defendants fleeing or the likelihood to re-offend. Increasingly, criminal justice systems use algorithms to support judges' decisions with machine predictions of the recidivism risk. These machine predictions are algorithms that derive their rules from data on past cases, corresponding information (like prior convictions) on the judges' decisions and information on recidivism (usually we allow up to two years after the exit from prison). To make human and machine decisions comparable, we evaluate the performance of algorithmic and human decision making on the same data.

There are various reasons why we would like to have a machine learning model to support decision making: judges are human and humans are naturally biased [13]. For instance, judge's decisions are influenced by very human factors such as hunger, or mood [8] or an unexpected loss of a prominent favourite football team [34]. Besides, it is easier to have algorithms adhere to strict processing rules, by programming them accordingly. However, we should be careful when considering the support of machine learning systems, since machine learning models inherit human bias (mostly through data), as mentioned above. Furthermore, technology is not value-neutral and it's created by people who express a set of values in the things they create. Often this has unintended consequences, such as discrimination or bias. Each data point in a dataset represents a criminal case which is characterised by a set of features. Machine learning models learn from these features to separate between recidivists and non-recidivists.

> "To make human and machine decisions comparable, we evaluate the performance of algorithmic and human decision making on the same data."

Most datasets on defendants in the criminal system contain features related to their demographics, and criminal history, e.g. age at main crime, sex, nationality, sentence, number of previous convictions, the year the crime(s) took place, whether probation was given or not, etc. These features are also known as static features because they were collected in the past and cannot be changed, even if the defendant would change her criminal behaviour.

Generally, a process or decision is considered fair if it does not discriminate against people on the basis of their membership to a protected group, such as sex or race. In this case, we tested for discrimination against for-

eigners and on the basis of sex. We detect discrimination in a decision making process (independent of a human or algorithmic origin) by testing the decision against a "ground truth" (e.g. recidivism occurred within two years after release date or not). Processes that are over proportionally correct or wrong for a particular protected feature indicate that there was discrimination.

SAVRY

Tools which assist decision makers are increasingly used in criminal justice, medicine, finance etc. One of the well-known tools for recidivism prediction is COMPAS, which stirred some controversy a while ago because arguably it discriminates against African-American people [1]. The risk assessment tool SAVRY is used in many countries across the world to assess the risk of violent recidivism amongst youth. It has been mainly designed to work well for violent crimes and male minors. In contrast to COMPAS, SAVRY is a transparent list of features/terms, and the final risk evaluation remains at the discretion of a human expert. SAVRY is based on a questionnaire which collects information on: early violence, self-harm, domestic violence, poor school achievement, stress and poor coping mechanisms, substance abuse, parent/caregiver with previous criminal convictions. Many of these features are known in criminology as dynamic features, because they can change over time as the defendant changes their behaviour.

What we do

The dataset we use in our experiments comprises of 855 offenders aged 12-17 in Catalonia of crimes committed between 2002-2010. The release year for all defendants in the sample was 2010 and the recidivism status was followed up in 2013 and 2015.

We train different machine learning models on three specific sets of features: (i) only non- SAVRY features (demographics and criminal history); (ii) only SAVRY features; and (iii) all features combined. We tested different machine learning methods and decided to take the top two performing methods to evaluate their fairness. We then compared these machine learning methods to the simple sum of SAVRY features (SAVRY sum) and the expert's assessment (Expert).

To evaluate predictive performance, we report the area under the receiver operating curve (AUC). This can take up a value between 0 and 1 and indicates how well a model is capable of distinguishing between positive and negative cases. The higher the AUC, the better the model is at predicting zeros as zeros and ones as ones. In decision situations like these, where there are only two potential outcomes (recidivism/no recidivism), there are two types of errors that can occur: (1) a defendant is wrongly classified as a high-risk type, although she would not have reoffended. We call this a false positive (FP) and this error contributes to the false positive rate (FPR); (2) a defendant is wrongly classified as low risk, although she has reoffended. We call this a false negative (FN) and this error contributes to the false negative rate (FNR). In Table 1 we illustrate how we compute these error rates, where Y represents the true outcome and $\hat{Y}$ represents the predicted classification.

		Predicted Classification $\hat{Y}=1$	Predicted Classification $\hat{Y}=0$	
True Outcome	Y=1	True Positives (TP)	False Negatives (FN)	False Negative Rate (FNR) FN/(TP+FN)
	Y=0	False Positives (FP)	True Negatives (TN)	False Positive Rate (FPR) FP/(FP+TN)
		False Omission Rate (FOR) FP/(TP+FP)	False Discovery Rate (FDR) FN/(FN+TN)	

Table 1: Computing the error rates in binary classification

To evaluate the fairness of a decision-making system (machine learning, SAVRY or the expert supported by

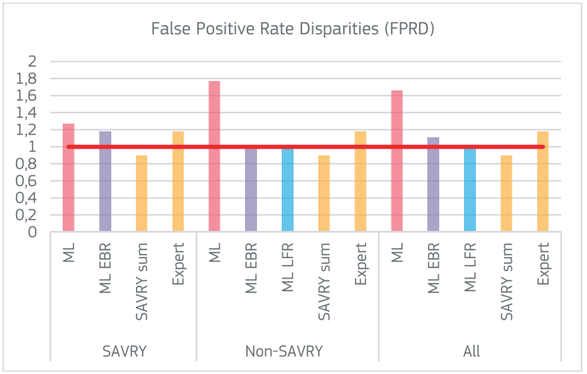
SAVRY), we compute these error rates separately, for foreigners and Spanish nationals and compute the error rate disparities, i.e. FPRD and FNRD by dividing the error rate for foreigners by the error rates for Spanish nationals. We say that a system is unfair if it commits relatively more errors for one group than for the other.

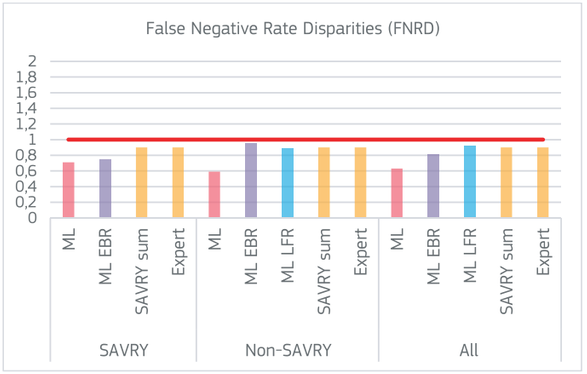
FPRD= {
- <1→ foreigners are less often wrongly classified as high risk as Spanish nationals
- =1→ foreigners and Spanish nationals are equally often wrongly classified as high risk
- >1→ foreigners are more oft en wrongly classfied as high risk as Spanish nati onals

FNRD= {
- <1→ foreigners are less often wrongly classified as low risk as Spanish nationals
- =1→ foreigners and Spanish nationals are equally often wrongly classified as low risk
- >1→ foreigners are more often wrongly classfied as high low as Spanish nationals

In our experiments the SAVRY Sum (AUC=0.64) and Expert (AUC=0.66) have a lower predictive power (lower AUC) than the ML models that achieve an AUC of 0.70 when trained only on non-SAVRY features and an AUC of 0.71 when trained on SAVRY and non-SAVRY features combined.

Figure 2: Disparity of error rates with respect to Spanish nationals. ML results are computed with logistic regression. Other ML techniques, such as multilayer perceptron yield similar results.





However, we also find that machine learning (ML) is generally less fair as it commits more decision errors for foreigners than for Spanish nationals. We see this in Figure 2, which illustrates the disparities of foreigners compared to Spanish nationals of the FPR in the top panel and the disparities of the FNR in the bottom panel for five different decision systems. Figure 2 shows, if we don't apply any correction to the ML algorithm (here we display the results of the ML method multi-layer-perceptron), ML commits more false positive errors and fewer false negative errors for foreigners compared to Spanish nationals. This disparity is much lower if the decision is made with only SAVRY or with a human expert.

In fact, the inclusion of non-SAVRY features (criminal history, age, sex, or nationality) increases the disparity between Spanish nationals and foreigners.

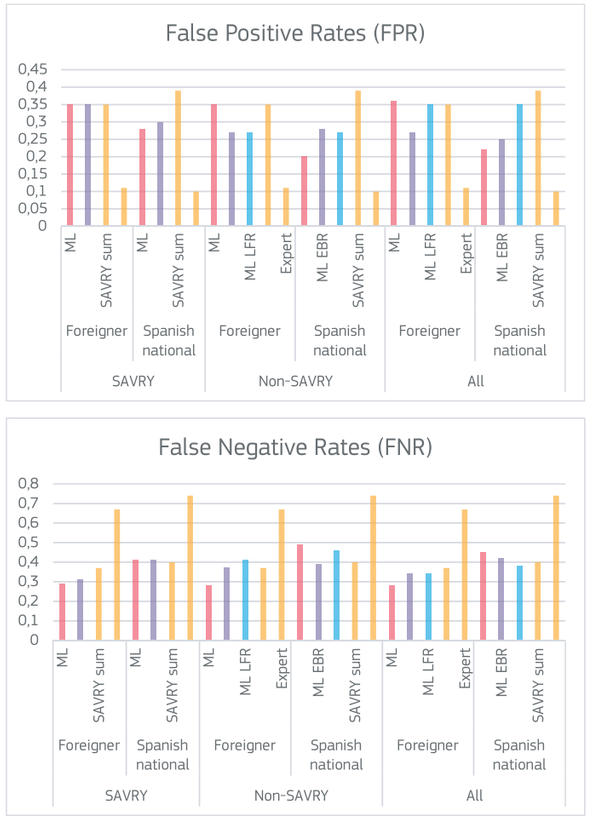


Figure 3: Comparison of FPR and FNR between Spanish nationals and foreigners. ML results are computed with logistic regression. Other ML techniques, such as multilayer perceptron yield similar results.

However, not using these non-SAVRY features is still discriminative, which can be seen in the first two bars of each panel. This was expected: many state of the art papers report discrimination even in the absence of sensitive features that are often correlated with the very group features that are tested for discrimination.

How well do algorithmic methods help mitigate algorithmic discrimination? We evaluate two methods that are meant to correct ML algorithms for fairness: (1) equalized base rates (EBR) and (2) learning fair representations (LFR) [30] and show the fairness results of both methods in Figure 2. EBR equalizes the prevalence of recidivism between foreigners and Spanish nationals in the data.

The LFR removes correlations of other features with the group feature for which we want to "remove" discrimination (in this case foreigner status). We can see in Figure 2 (purple and blue bars) that both methods reduce disparities in terms of both error rates. In most cases the correction methods move the error rate disparities closer to 1, where error rates occur equally for both Spanish nationals and foreigners. However, at what cost do we achieve this equalisation of error rates? To better understand how EBR and LFR work, we plot the FPR and FNR for Spanish nationals and foreigners instead of the disparity between the two groups in Figure 3. Again, the results of the corrected algorithms are represented with purple (EBR) or blue (LFR) bars.

We find that EBR and LFR, in order to reduce disparity in error rates, increase the FPR for the group which has been initially positively discriminated (Spanish nationals) and increase the FNR for the group that was initially discriminated (foreigners). So in the end a reduction in the error rate disparity is achieved only by increasing the actual error rates (and consequently increasing individual or public harms) for specific groups. Interestingly we find that expert evaluations tend to have, for all groups, a substantially lower FPR and a substantially higher FNR than the other decision systems. This suggests that in the context of juvenile justice, case workers tend to follow a more lenient decision policy towards the defendants. That is, case workers are more inclined to "let a high risk candidate go", than to falsely accuse a defendant of being a high risk candidate. This example illustrates how normative values of "what is fair" are reflected in decision making.

"Case workers are more inclined to "let a high risk candidate go", than to falsely accuse a defendant of being a high risk candidate. This example illustrates how normative values of "what is fair" are reflected in decision making."

Moreover, the difference between the three sets of features led us to analyse the importance of the features by using machine learning interpretability. With respect to that, while some models are interpretable by definition, for other black-box models we need post-hoc methods which can give an approximate interpretation. One of these frameworks, possibly the most used is LIME [22]. Applying LIME to the multi-layer perceptron prediction shows that the model relies more on the non-SAVRY features (sex, age, nationality), rather than SAVRY features. This suggests that the machine learning models in this context assess the recidivism risk of defendants mostly based on features that defendants could not change. In contrast, the expert's assessment relies more on the SAVRY features for which the defendant has opportunities to change them. This is another issue that makes risk assessment that is only based on machine learning problematic.

Our conclusions

As we mentioned, not using static (or especially demographic) features does not prevent discriminating outcomes. This made us think that the causes of discrimination do not lie in the features but in the data itself. It is established in the literature that one of the important sources of discrimination is the data itself. A good indicator of this is the difference in the prevalence of recidivism, also termed "base rates", between Spanish nationals and foreigners. Basically, if 46% of the foreigners in your dataset are recidivists compared to solely 32% of Spanish nationals, then your model will learn that the features correlated to being a foreigner are important. Such a model will yield more false alarms with respect to foreigners. We also showed that processing the data before training the algorithm in a way that removes the importance of sensitive features (in this case foreigner status) does not guarantee us better outcomes.

Finally, we highlight that processing the data that lead to discriminatory outcomes does not solve the discriminatory issues (such as over-policing of foreigners) that led to biased data in the first place. In fact, such mitigation measures may exacerbate discrimination in criminal procedures by hiding the discriminatory steps that occurred before the data was created and perpetuating the problem through harmful decisions. Therefore, understanding the causes of recidivism and taking into account the feedback of the implementation of the algorithm in the real world is more beneficial than brute-forcing the decision-making system to fit certain fairness metrics. Therefore, we conclude that algorithms that predict recidivism in criminal justice should only be used with caution, under strict protocols and under the awareness of such fairness issues.

> "Causes of discrimination do not lie in the features but in the data itself."

The occupational impact of AI

AI is poised to have a transformative effect on almost every aspect of our lives, from the viewpoint of individuals, groups, companies and governments. While there are certainly many obstacles to overcome, AI has the potential to empower our daily lives in the immediate future. In this regard, the impact on the labour market is already visible, but the workplace may be transformed in the following years. While past technologies could only automate tasks that follow explicit, codifiable rules, AI and machine learning technologies (ML) can infer rules automatically from the observation of inputs and corresponding outputs. This implies that ML may facilitate the automation of many more types of tasks that were not feasible previously. However, there is a high degree of uncertainty when it comes to determining whether a problem can be solved, or an occupation or task can be replaced or automated by AI today [18].

What we do

In this context, we develop a framework for analysing the occupational impact of AI progress. The explicit focus on AI distinguishes this analysis from studies on robotisation, digitalisation and online platforms and the general occupational impact of technological progress. The framework links tasks to cognitive abilities, and these to indicators that measure performance in different AI fields.

More precisely, we map 59 generic tasks from the worker surveys European Working Conditions Survey (EWCS), the Survey of Adult Skills (PIAAC), as well as the occupational database O*Net to a set of intermediate layers of 14 cognitive abilities (see [31] for a further description), including: memory processes (MP); sensorimotor interaction (SI); visual processing (VP); auditory processing (AP); attention and search (AS); planning, sequential decision-making and acting (PA); comprehension and expression (CE); communication (CO); emotion and self-control (EC); navigation (NV); conceptualisation,

learning and abstraction (CL); quantitative and logical reasoning (QL); mind modelling and social interaction (MS); and metacognition and confidence assessment (MC).

Regarding the aforementioned abilities, we integrate several theories of intelligence and cognition in psychology, animal cognition and AI literature to give a broader definition of what are cognitive abilities, as a more independent latent layer (latent being something which is inferred from human behavior rather than being directly observed from it) than human abilities (work-oriented) or AI abilities (technology-oriented). We finally map these cognitive abilities to a comprehensive set of around 350 AI benchmarks, competitions and tasks (which are metrics on publicly available datasets that indicate progress in AI techniques) based on previous analyses and annotations of AI papers [11, 14, 15] as well as open resources such as *AICollaboratory*[2] [16, 13], thus ensuring a broad coverage of AI tasks (see Figure 4).

Cognitive abilities are therefore a better parameter to evaluate progress in AI [12]. We focus on abilities instead of skills because, from a human perspective, abilities are innate and primary, whereas skills are acquired through a combination of abilities, experience and knowledge [9]. Since knowledge and experience are not appropriate properties of AI, linking AI benchmarks to abilities (instead of skills) should be less prone to measurement error [12].

This cognitive ability perspective allows us to distinguish machines that, through AI, are empowered with the abilities of performing in a range of several tasks from those machines that are constructed or programmed solely to perform a specific task. For instance, the ability of understanding human language (covered by the area of Natural Language Processing) can be applied to a variety of tasks (such as reading or writing e-mails or advising clients).

2 http://www.aicollaboratory.org/

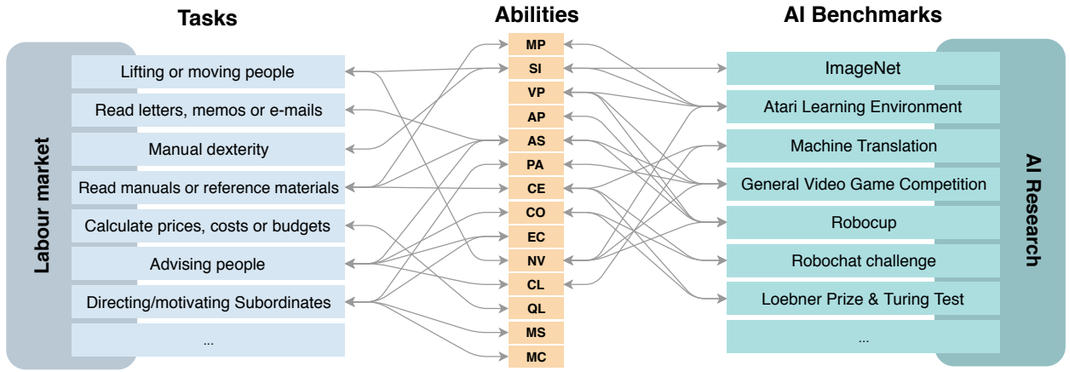


Figure 4: Bi-directional and indirect mapping between the job market and Artificial Intelligence

Our findings

Through this approach, we may gain a broader understanding of the occupational impact of AI (e.g. over tasks, occupations, sectors, countries, etc.). That is, the framework allows us to identify which abilities are less likely to be performed by AI and are therefore less prone to changes in how they are currently being performed. In this regard, we have observed that most of AI exposure is driven by its impact on tasks that require intellectual abilities, such as comprehension, attention and search. High-skill occupations such as medical doctors and teachers are thus more exposed to AI progress than comparatively low-skill occupations such as cleaners, waiters or shop salespersons.

> “High-skill occupations are more exposed to AI progress than comparatively low-skill occupations.”

Furthermore, the framework also allows us to examine the relationship between the distribution of research intensity (or activity) in AI research (e.g. where the AI research community is putting the focus) and the relevance for a range of work tasks (e.g. what

areas of AI research activity would be responsible for a desired or undesired effect on specific labour occupations) in current and simulated scenarios[3] [18].

In this regard, we have found that some of the least used abilities in labour tasks are precisely those where more progress is apparently taking place in AI, such as visual and auditory perception using deep learning (e.g., for facial recognition, image generation, speech recognition, etc.); and sensorimotor interaction, through (deep) reinforcement learning (e.g. for game playing, robotics, etc.). There are many reasons for this finding: (i) some of these abilities are taken for granted, or they could be covered by more rudimentary sensoric abilities and are therefore not mentioned explicitly (e.g. recognising objects and moving around in the workplace); (ii) some of the tasks in the workplace require skills for which there is not a high AI research activity at the moment. Regarding the latter, it seems that, for instance, there is not much interest in having AI perform some tasks that require social interaction, at least not before more socially isolatable tasks can be fully performed by AI.

> **“We can be much more certain about the capacity of AI to transform jobs than about its capacity to destroy them.”**

Overall, this framework presents an appropriate way to measure the relationship between AI progress and labour markets. Cognitive abilities, as the link between these two entities, can capture well the general advances in data collection for both labour markets and AI research. It should be noted, however, that in this framework AI exposure does not necessarily mean automation. The previous findings do not imply that purely intellectual tasks will be automated, as other processes could occur when technology takes over some work that was previously performed by a human. In the end, high AI impact in a task could imply that the way a task is being performed is just restructured. Moreover, we find that to most occupations social abilities are highly relevant, while AI progress has a stronger effect on abilities that deal with intellectual tasks. Corresponding labour market processes could potentially increase the demand for workers with strong social abilities. Overall, we can be much more certain about the capacity of AI to transform jobs than about its capacity to destroy them.

3 We develop an online visual approach for showing the intensity flows between AI benchmarks and labour market tasks and occupations: https://safe-tools.dsic.upv.es/shiny/OTAAI/

Social robots and human development

In this scenario, we examine the impact of Embodied AI (robots) on human cognitive and socio-emotional behaviour. Taking into consideration research in human psychology and the role of embodiment in human perception and development [29], we hypothesise that the embodied and physical nature of robots affects human behaviour in ways that might be distinct from non-embodied intelligent systems. To explore the role of robots on human development, we focus on two scenarios. First, we focus on the human problem-solving process and the ways that specific robot task-related interventions affect human cognitive mechanisms and social interactions. Second, we examine the ways in which human trust develops during the interaction with a social robot, by manipulating the robot's behaviour in terms of social features. Our research contributes to with novel findings to the scientific field of human-robot interaction and informs policy-related discussions on embodied AI and children's rights.

What we do

To address the above-mentioned goals, we design and conduct behavioural experiments in human-robot interaction (HRI) settings. To conduct these HRI experiments, we design specific robot behaviours that serve our research goals and hypotheses, which are then embedded in specific real-life scenarios, such as in school-settings. Our findings aim to contribute to the scientific dialogue about human-centered robot behaviour design. At the same time, during the course of our research we are aware of emerging ethical considerations in relation to human rights. In the following paragraphs, we first briefly describe a set of human-robot interaction experiments that have been conducted in the context of the HUMAINT project. Then, we discuss our approaches in possible emerging risks in the case of child-robot interaction and we present a set of scientific outreach activities relevant to this topic.

For the project, we have conducted a set of human-robot interaction

experiments on two main topics: (i) robot-assisted problem-solving with child users; and (ii) development of trust in human-robot interaction. We selected one of our target groups to be children because their cognitive, behavioural and socio-emotional skills are under rapid development. This means that the integration of robotic technologies in their everyday life might impact their development in unique ways, which are yet to be systematically explored. Additionally, children are expected to be prepared in the most appropriate and effective way not only as current and future users of robotic systems but as critical thinkers and as potential actors to be actively involved in the design and development of those systems.

All the empirical studies of human-robot interaction were reviewed and approved by an ethical committee which was created ad hoc at DG JRC. The ethical committee consisted of internal and external experts on experimental studies with human subjects. Written informed consent to participate in these studies was provided by all participants or the participants' legal guardian. In the case of child-robot interaction studies, all children assented to participate.

Robot-assisted problem-solving with child users

Child-robot interaction is a field that aims to understand how children develop when interacting with robots. The community of child-robot interaction is fast growing and there are already some initial results in several contexts. Our experiment aimed to understand the development of children's problem solving in a controlled setting. We considered the paradigm of the Tower of Hanoi (Figure 6), which is a task of incremental difficulty and we conducted an HRI behavioral experiment to evaluate task performance. We used the Haru robot for which we designed specific behaviour for verbal interaction (see Figure 5). The Haru robot is a tabletop research robotic platform that is still under development by the HONDA Research Institute, Japan. It presents different modalities for actuation and it can move in 5 degrees of freedom (base, neck, eyes tilt, eyes roll, eyes stroke).

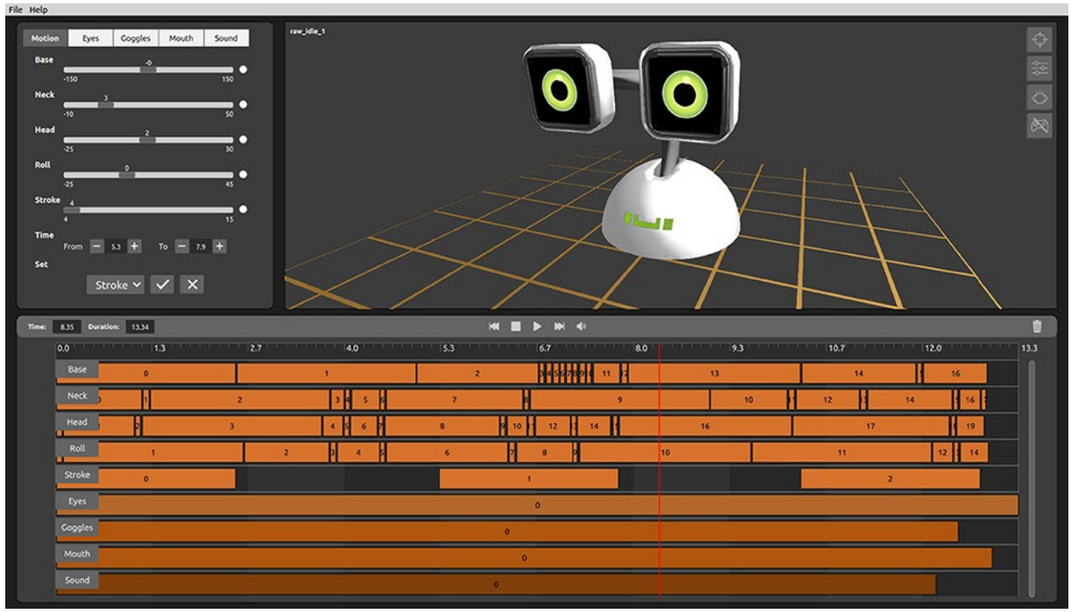


Figure 5: Snapshot from the interface for the Haru robot behaviour design. Figure reproduced from [4]

For our experiment, we designed two types of robot interventions, "voluntary" and "turn-taking", manipulating exclusively the timing of the intervention. Twenty primary school children participated in the experiment. In the turn-taking condition, the child and the robot played in turns trying to solve the problem together. The robot was programmed in such a way that it always gave the optimal suggestion. In the voluntary interaction condition, the child was invited to solve the problem by themselves and they were free to ask the help of the robot whenever they wanted. During the voluntary interaction condition, the robot was present and constantly aware of the current status of the game.

Our results indicate that the children who participated in the voluntary interaction setting showed a better performance in the problem-solving activity during the evaluation session despite the large variability in the frequency of self-initiated interactions with the robot. We note that the sample of this study was relatively small, which didn't allowed the use of parametric statistical tests. The use of the non-parametric Mann–Whiteney's U-test, which revealed a statistically significant difference in the performance between the two conditions during the evaluation phase ($p = 0.038$, $a = 0.05$). Additionally, we present a detailed description of the problem-solving trajectory for a representative single case study, which reveals specific developmental patterns in the context of the specific task.

The findings of this study indicate that when designing intelligent robotic systems that aim to assist children in problem solving, it is important to allow children to explore and to initiate the interaction with the robot according to their needs. This is in accordance to the principles of child-centred design of robots.

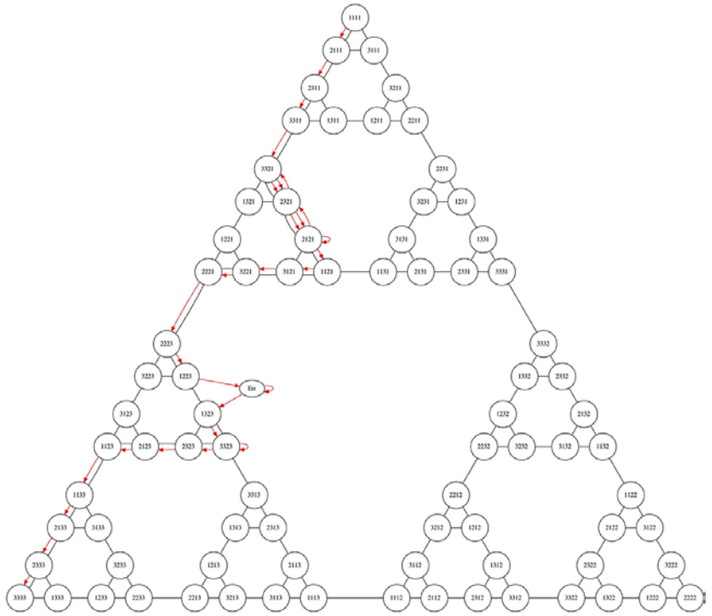


Figure 6: Graphical representation of a child's task performance of the Tower of Hanoi. It involves three vertical pegs and a fixed number of coloured disks with graduated sizes that fit on the pegs. At the outset, all the disks are arranged on one of the pegs like a pyramid with the largest disk on the bottom. It requires the arrangement of disks from an initial starting point to a specified end point in the minimum number of moves according to some predefined rules. Figure reproduced from [4]

> "... it is important to allow children to explore and to initiate the interaction with the robot according to their needs. This is a principle of child-centred design of robots."

Figure 7: Triadic interaction of two children and the Huru robot in a study-setting held at the Colegio Internacional de Sevilla San Francisco de Paula

Development of trust in human-robot interaction

The psychological construct of trust is among those which have a special interest in HRI research. However, it is not clear yet how various robot design features in terms of verbal and non-verbal communication can affect how humans perceive the robots' performance and elements of social expressivity as indicators of their trustworthiness. Furthermore, little is known regarding the human developmental of trust over time and how this is affected by the robot's previous behaviour and possible violation of trust.

For the investigation of elements of human trust towards robots, we conducted two case studies. In the first one, we built upon the study described in [4] and we designed a behavioural experiment with a triadic interaction: a pair of children interacting with the Haru robot with the aim to solve the Tower of Hanoi task.

We recruited 82 primary school children. We first invited the children to complete a trust belief validated questionnaire which indicated the levels of trust belief in the context of children's existing experiences in the school environment. Then, the children interacted with the robot to solve the Tower of Hanoi. During the intervention, we manipulated the robot's behaviour in terms of its cognitive reliability and features of its social behaviour [7].

Figure 8: The PEPPER robot (Softbanks Robotics)

Social robots and human development

After the intervention, the children solved an increased difficulty level of the Tower of Hanoi with voluntary interaction, being free to request the help of the robot. Lastly, we conducted a post-intervention semi-structured interview with all the children separately to research their perceptions about the robot. The analysis of the data of this study is still ongoing. Initial results indicate an effect of the robot behaviour on child to child social dynamics. More specifically, we observed that the children who participated in the intervention with a non-reliable robot tended to exhibit more collaborative behaviour during the session than children who participated in the intervention with the reliable robot. This indication can inform robot-design behaviours that can support child-to-child collaboration collaboration and social dynamics.

In the second case-study, we focused on adults' development of trust towards robots with regards to their perceived vulnerability to a robot. We applied the situation in three every-day scenarios. In each of the scenarios we designed an instance of trust violation in one of the following ways: (i) by hypothetically misusing the user's private information; (ii) by being inconsistent in terms of hypothetical financial exchanges with the user; and (iii) by expressing non-logical assumptions during the communication with the user without being transparent.

"Initial analysis, however, indicates the importance of contextualization when researching trust in human-robot interaction."

We first used a validated Multi-Dimensional-Measure of Trust (MDMT) questionnaire to map our participants' existing perceptions and preconceptions about robots in various settings. Then, the participant interacted with the robot in one of the three scenarios of the trust violation. At a post-intervention session, we used a part of the MDMT questionnaire [28] to measure human-robot trust. Lastly, we conducted semi-structured interviews to understand the human perceived vulnerability in the context of the experienced robot trust violation. We first conducted a technical pilot study with the PEPPER robot (Figure 8) to test the design elements of these experiments as well as the robot's behaviours; then, we conducted an online survey with 96 participants who went through all the phases of the experiment and took part in a screen-based interactive robot scenario. A physical HRI experiment with the same design and assessment instrument has been planned and it will provide us with further insights regarding the role of embodiment on human-robot trust. Therefore, analysis of our findings is still ongoing. Initial analysis, however, indicates the importance of contextualization when researching trust in human-robot interaction.

Bio-inspired robot behaviour

In a separate experiment, we replicated the experiment of child-robot problem-solving, this time only with two virtual robots to see if the virtual robot behaves or learns in a similar way as a child does. With our HRI experiments, we aim to understand the human behaviour in the context of human interaction with robots. However, insights of HRI research might contribute to the research and development of autonomous robotic systems that might learn in similar ways as humans learn. Transferring mechanisms of the human brain to AI is an ambitious yet promising goal for the advances of the state of the art in AI and robotics. Towards this aim we started from hypotheses derived from our empirical study [4] and we tried to examine if they are validated in the same way for a basic reinforcement learning algorithm. Thus, we replicated the child-robot interaction experiment and checked whether receiving help from an expert when solving a simple close-ended task (the Tower of Hanoi) allows to accelerate or not the learning of this task, depending on whether the intervention is taken in turns or requested by the player [2]. Our experiments with two autonomous agents have allowed us to conclude that, whether requested or not, a Q-learning algorithm benefits in the same way from expert help as children do. In a similar line, we have identified elements of psychological mechanisms in other contexts such as in child's music-making [3], which can be used for the design of robot-assisted music-making for children.

AI for music creation and listening

The application of AI to music has already been studied in the literature for many decades and presents numerous unique opportunities for a variety of uses, such as the recommendation of recorded music from massive commercial archives or the (semi-)automated creation of music. In these contexts, AI can produce outcomes in a domain fully entrenched in human creativity. Various participants contribute to and benefit from music including writers, composers, producers, musicians, educators, listeners and music organisations. Data-driven machine learning techniques are currently exploited to generate music in a semi-automatic way and to provide music recommendation and retrieval in large data collections.

In HUMAINT, we address the social, cultural and ethical issues related to the use of machine learning in the musical context. In [24], we address two independent perspectives of AI applied to music creation: copyright law and engineering practice, and we discuss a set of questions related to these two views. In terms of copyright law, a major requirement to establish authorship recognition is to establish who is accountable for music-AI systems and how. With respect to "who", we need to consider traditional stakeholders involved in the music creation process, e.g. the writer, composer, singer, musicians, and music producer, but also new ones brought by the use of AI techniques such as the creators and curators of the music collections (datasets) used for training machine learning models, the developers of the AI system, and their users (e.g. composers, producers). In order to establish "how", i.e. the contributions of each stakeholder to the final musical piece, we need to define ways to document in a transparent way and with a suitable level of detail the working principles of the AI systems and the degree to which AI is involved in the process. Transparency and accountability have other benefits, e.g. to inform listeners about the use of AI in the music they hear or to study the way AI impacts music-related jobs as we have seen in section 4. In addition to authorship,

artistic value is of major relevance here, it depends on people's perception, and it remains an open question for future research.

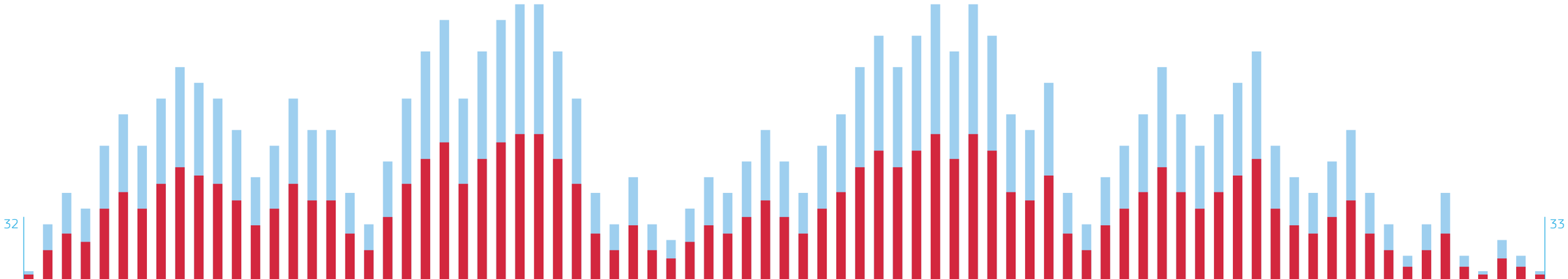
The origins of Music Information Retrieval (MIR) technologies can be traced back to the late 90's, thanks to the internet, audio compression techniques, increasing computing power, and the appearance of music players and streaming services for unlimited music consumption, anytime and everywhere. The goal of MIR technologies is to help users find music in large repositories, and they are usually based on the concept of music similarity. Over the last years, data-driven machine learning techniques have become predominant in music-recommender systems, and their design impacts the music we listen to in streaming services. We focus our study on the analysis of the diversity of these recommendations. In [21] we define a set of metrics for the statistical modeling of popularity and semantic diversity in music playlists, and we apply it to more than 400 000 playlists from four datasets, created in different temporal and technological contexts. We observe how differences between datasets emerge, reflecting the context in which playlists have been created. We find it extremely valuable to compare different playlist datasets, as it allows us to understand how changes in the listening experience are affecting playlist creation strategies: from radio stations, user-generated to algorithm-generated playlists.

In [20], we present a framework to assess the impact of music recommendation diversity, or the lack thereof, on music listening experiences. This involves: (i) the formalisation of a working definition of diversity in the music field; (ii) the development of evaluation practices for diversity in the context of music-recommender systems; (iii) the observation of the emerging impact due to music recommendation diversity; and (iv) the proposal of countermeasures for mitigating negative or reinforcing the positive impact observed.

AI for music creation and listening

"AI is changing and can change musical taste, opinion about music and culture, and our relationship with music for better and for worse."

In this scenario, HUMAINT contributes to understand how AI is changing and can change musical taste, opinion about music and culture, and our relationship with music for better and for worse.



AI in medicine and healthcare

The use of machine learning approaches to target clinical problems is expected to change clinical decision-making. Our research tries to understand this change. In [23] we analyse the classical or traditional pathway by which clinicians make decisions and study how machine learning can have an impact in this pathway at four different levels, as explained in Figure 9.

By observing the current status of AI technologies, we identify the current clinical status and challenges associated with each of these levels identified in Figure 9, together with the challenges related to the machine learning process. We look at data acquisition, predominantly by extracting standardised, high-quality information with the smallest possible learning curve; at feature extraction, by discharging healthcare practitioners from performing tedious measurements on raw data; at interpretation, by digesting complex, heterogeneous data in order to improve the understanding of the patient status; and at decision support, by leveraging the previous step to predict clinical outcomes, the best response to treatment or to recommend a specific intervention.

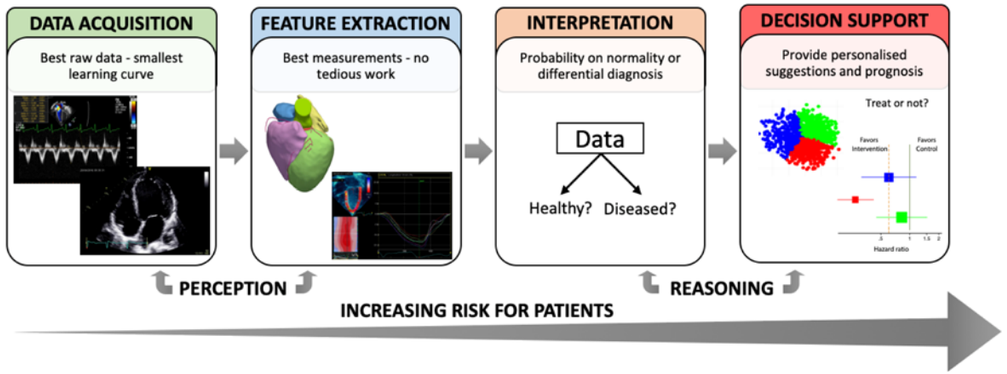


Figure 9: Machine learning in the clinical decision-making pathway. Figure reproduced from [23]

Apart from clinical decision making, there are other areas in medicine and healthcare where AI is having an impact. We present in [11] a literature review of around 600 references, including scientific publications, product descriptions and press articles on the use of AI in medicine and healthcare. This review studies the state of the art of research and technology, including software, personal monitoring devices, genetic tests and editing tools, personalised digital models, online platforms, augmented reality devices, and surgical and companion robotics. We then classify these applications in terms of their technology availability level (TAL, which is a qualitative indicator of how available a technology is in real-world scenarios) and the potential risks they can have for human welfare.

Our findings

When we study the use of machine learning in the clinical decision-making pathway, we observe an increasing risk for patients in the use of machine learning methods, as illustrated in Figure 9. Privacy of the data used in machine learning becomes critical in the medical context due to the sensitivity of health records. In addition, transparency and explain-ability (i.e. the ability to explain the internal mechanics of a machine learning system in human terms) of algorithms are a major technical requirement in the medical domain together with auditability (i.e. the possibility to carry out a thorough evaluation) and traceability (i.e. keeping and exhaustive record of all algorithmic processes), as it is crucial to understand and document the working principles of systems for its trustworthy adoption. We conclude that machine learning research in the clinical decision-making context should advance beyond the learning of single, repetitive tasks to systems that are aware of the full context of clinical decision making. The above requires

“Privacy of the data used in machine learning becomes critical in the medical context due to the sensitivity of health records.”

multidisciplinary teams analysing the processes currently in use, algorithms that can consider longitudinal data and integrate heterogeneous information sources, and good strategies for human-algorithm interaction and trust.

From our analysis of the TAL of AI systems exploited more broadly in medicine and healthcare, we observe that new paradigms arise, such as the concept of extended personalised medicine (defined as the exploitation of heterogeneous data sources such as social media data, environmental factors or medical records, for tailored diagnosis and treatment) and a change in the public perception of medical AI systems, and how they show, simultaneously, extraordinary opportunities and risks. In addition, we witness the transformation of the roles of doctors and patients in an age of ubiquitous information and identify three main paradigms in AI-supported medicine: fake-based, patient-generated, and scientifically tailored views. We conclude that in medicine there are some specific policy challenges, including the need of informed citizens and new social and ethical aspects to consider (e.g. related to the fundamentals and definitions of life).

Our work in this area is specially related to the exploitation of AI systems to fight the COVID-19 disease. We discuss in [33, section 3] some relevant aspects related to the impact of AI in human behaviour in the context of the current pandemic. These include the boost of telemedicine due to the limitation of physical contact, the benefits and risks of data-driven algorithms for COVID-19 diagnosis, epidemiological prediction and clinical management, the balance between individual rights and public health, and the change in public perception of robots, i.e. from fear to acceptance and new roles they are adopting in this new scenario to replace humans in situations which are not risky, such as measuring patients' temperatures.

Diversity in AI

During its research activities, the HUMAINT project has confirmed that there is a need for diverse working teams, including varied disciplines and gender diversity, in the development of AI systems to reduce bias in technologies/algorithms and to ensure that they are meaningful for everyone. From this finding, HUMAINT has incorporated the diversity dimension as transversal to its research activities. We present here the two main aspects we have been researching on: diversity in criminal justice and of major AI conferences.

As presented in detail in section 3, we looked at bias and fairness in algorithmic decision making in criminal justice, including at the impact of sex, highlighting gender bias. For instance, in [26], we compare two different strategies (human versus machine) to assess the risk of recidivism in juvenile criminal justice. In our sample of 855 minor defendants in Catalonia who were all released in 2010, we find that about 40% of males are re-offending compared to 20% of females. This is a known fact among criminological experts and case workers, and they take this into account when assessing the risk of females re-offending. However, this could become a problem if the risk assessment is being done by a machine. If the gender variable is not explicitly considered, it means that the machine will treat males and females equally in their risk assessment. Because females are the minority in such datasets, the algorithm would be biased towards the behaviour of males. This would lead to higher likelihood among females to be falsely accused of having higher risk of reoffending. If case workers act on these algorithmic recommendations, this could seriously harm the prospects of females in criminal justice.

The second aspect we address is the diversity of major AI conferences, reflecting researchers and developers of AI systems. In order to implement and assess the impact of policies to achieve diversity, it is essential to develop meaningful diversity indicators. Our project promotes the divinAI initiative[4], which develops a set of indicators of the heterogeneity and diversity of AI conferences' committees, authors and keynote speakers, where we try to monitor diversity in terms of gender, academia vs. industry, and geographical location [10]. divinAI has organised several 'Hackfests' on the topic to crowd-source conference data while raising awareness on this problem.

As an example, Figure 10 illustrates the results for the International Conference on Machine Learning (ICML) 2019. The gender diversity index for ICML 2019 is 0.82 (an index of 1 means an equal distribution of men and women). Considering that women are a minority in STEM (Science, Technology, Engineering and Mathematics), the results are in line with expectations. However, if we break down the results, between author submissions, invited keynote lecturers, and organisers, we notice that women only have a share larger than 50% among keynote speakers. Compared to this, about 90% of all authors are male and among organisers 60% are male[5].

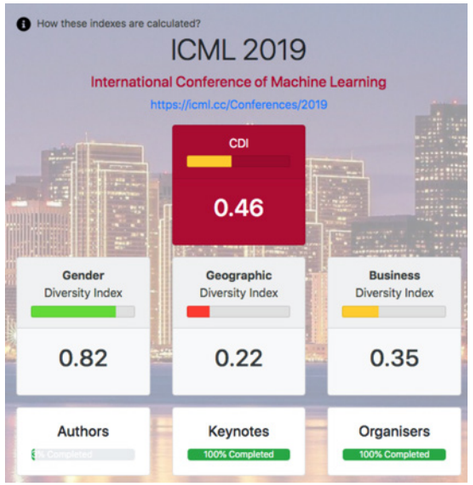


Figure 10: Diversity indexes for the International Conference on Machine Learning 2019

4 https://divinai.org

5 The number of authors must be considered with caution because only a small share of all authors is evaluated.

Emergent ethical considerations and community building

Throughout the course of our research, we have been constantly reflecting on the emerging ethical considerations in our work and building interdisciplinary communities around our research topics. As part of the HUMAINT project, we have organised a yearly open event, the HUMAINT Winter School, having its 1st edition on the topic of AI and its ethical, legal, social and economic (ELSE) impact (February 4-8th, 2019, Seville, Spain[6] and its 2nd edition on the topics of fairness, accountability and transparency of AI (January 22-24th, 2020, Seville, Spain[7]). The annual winter school is conceived as an interdisciplinary course about AI systems and has contributed to community building and to create open and reproducible teaching materials around our topics of interest (lecture slides, videos and other resources).

In the scope of our work on criminal risk assessments, we have contributed to an expert meeting on predictive policing of the Police and Human Rights Programme (PHRP) of Amnesty International, The Netherlands. The meeting brought together representatives of civil society organisations, law enforcement agencies, international institutions, academic institutions and industry to discuss the increasing use of algorithms in policing and its consequences for human rights[8].

In the context of human-robot interaction studies and with special focus on the child user, we have contributed to the identification of possible risks in child-robot interaction by initiating discussions with relevant stakeholders in the form of a series of relevant workshops in conjunction with high impact conferences. First, we discussed about the current state of the art in child-robot interaction by identifying the challenges and the emergent opportunities [3]. Then, we discussed the role of robot expressivity in human-robot interaction [7] and we established a theoretical model for cross-cultural research, as one of the steps of inclusion in HRI research, regarding the occurrence of low-level non-verbal social features [5]. In the same context we contributed to the kick-off discussion on AI and children's rights initiated by UNICEF, New York[9]. Lastly, according to our vision for children's awareness, inclusion and participation, we initiated a series of workshops on identifying methods to empower children's reflections on AI, Robotics and other intelligent technologies [6]. Our future research questions will focus on challenges as well opportunities we foresee to emerge in children's interaction with robots in various settings.

6 https://ec.europa.eu/jrc/communities/en/community/humaint/event/ humaint-winter-school-ai-ethical-social-legal-and-economic-impact

7 https://ec.europa.eu/jrc/communities/en/community/humaint/event/2nd-humaint-winter-school-fairness-accountability-and-transparency

8 https://www.amnesty.nl/actueel/phrp-expert-meeting-on-predictive-policing

9 https://ai4children.splashthat.com/

Conclusions

The HUMAINT project researches on the impact of AI systems on human behaviour in different contexts, and over a short period of time. From our work, we identify four main challenges we are currently addressing and recommendations for future research.

First, there is a limitation in accessing open and representative behavioural data needed to study human-AI interaction, especially in sensitive and complex scenarios. Currently, it is difficult for researchers to access large-scale and representative repositories, particularly in academic contexts where research is disconnected from real-world AI-powered products. This limits the use of representative data in existing studies and introduces a bias in the literature that is directed towards industry contexts. We consider behavioural data collection as a key sensitive matter as, on the one hand, it is a source of knowledge for research but, on the other hand, needs to be carefully aligned with data protection guidelines. Our project combines the use of existing open data [26] complemented by data collection through behavioural studies (e.g. in child-robot interaction). One future scenario is to design trustworthy[10] user-centred platforms for research, to strengthen the collaboration between industry and academia for a more comprehensive research on human-AI interaction, and to formalise responsible and reproducible behavioural research and protocols.

10 https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

A second challenge we identified is the definition of standard methodologies and appropriate metrics for the evaluation of AI systems (e.g. performance-based measures as well as others focused on further relevant aspects such as robustness, transparency and fairness) and for running user studies in human-AI interaction. The lack of common methods makes it difficult to reproduce, contrast and build on top of existing studies. Our project contributes to open methodologies and protocols for the scientific community to carry out behavioural and data-driven analyses. This includes, for instance, the HUMAINT repository for fairness analysis of machine learning decision making[11], the AI collaboratory for the evaluation, comparison and classification of AI[12] and the divinAI indicators on the diversity of AI events[13].

A third challenge we address is the need to carry out this research in the intersections of different disciplines, with diverse teams (e.g. in terms of gender or cultural background) and where AI system developers must understand the social context where AI systems are embedded and AI users must understand the working principles and limitations of the systems they are using. This is one of the challenges of the work we have presented here, as we integrate knowledge from economics, robotics and machine learning, music creation, policy-making and education.

The final challenge involves existing literature and studies focusing primarily on assessing impact in the short term. We see the need to incorporate mid-to-long-term studies and literature as well as foresight studies into the current literature for a more comprehensive understanding on how AI is and will change human behaviour.

Our work has brought to light the amazing opportunities and potential of AI for the betterment of human welfare, wellbeing and for society as a whole. However, as we have highlighted, such a technology is not without its limitations and we recognise that our progress with AI is only at the beginning stages. With this in mind, we continue our research with full awareness of AI's powers and perils.

11 https://gitlab.com/HUMAINT/humaint-fatml
12 www.aicollaboratory.org
13 www.divinAI.org

Who are HUMAINT?

We are an interdisciplinary team of researchers from engineering, computer science, cognitive science and economics. Our motivation is to research on the social, economic, ethical and cultural impact of Artificial Intelligence.

- **Emilia Gómez** (lead scientist) has a research background in music information retrieval. Starting from the music domain, she researchers on the impact of AI systems on human behaviour, including the topics addressed in this document.
- **Vicky Charisi** conducts research at the intersection of embodied AI (robotics) and human behaviour with a focus on child-robot interaction and examines the impact of intelligent systems on human cognitive and socio-emotional development.
- **Marius Miron**'s research interests is on Explainability and Fairness in Machine Learning.
- **Songül Tolan** conducts research on the impact of AI on labour markets, data-driven decision making and the evaluation of machine predictions against human decision-making in terms of Fairness, Accountability and Transparency.
- **Fernando Martínez-Plumed**'s main research and academic interest encompasses several areas of Artificial Intelligence, machine learning, IRT, inductive programming, cognitive systems, data science and visualisation.
- **Marina Escobar** is a developer with the HUMAINT project.

The HUMAINT project has also involved other researchers from different institutions, including:

- **Dr. Carlos Castillo,** Universitat Pompeu Fabra, Barcelona, Spain.
- **Prof. Virginia Dignum,** Umea University, Sweden.
- **Dr. Ana Freire,** Universitat Pompeu Fabra, Barcelona, Spain.
- **Prof. Emilio Gómez-González,** Group of Interdisciplinary Physics, Escuela Técnica Superior de Ingeniería (ETSI), Universidad de Sevilla, Spain.
- **Ms. Glenda Hannibal,** TU Wien, Austria.
- **Prof. José Hernández-Orallo,** Universidad Politécnica de Valencia, Spain.
- **Dr. Michael Mathioudakis,** University of Helsinki, Finland.
- **Dr. Luis Merino** and **Dr. Fernando Caballero,** Universidad Pablo Olavide, Seville, Spain.
- **Dr. Pablo Negri,** Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Argentina.
- **Dr. Sergio Sánchez-Martínez,** IDIBAPS, Barcelona, Spain.
- **Dr. Bob Sturm,** Royal Institute of Technology, Sweden.
- **Members of the Music Information Research Lab,** Music Technology Group, Barcelona, Spain.
- **Colegio International de Sevilla** – San Francisco de Paula, Seville, Spain.

Finally, HUMAINT has collaborated with other research programmes and initiatives in related topics, mostly:

- AI WATCH, the EC knowledge service to monitor the development, uptake and impact of AI in Europe.
- The International Consortium for Socially Intelligent Robotics.
- TROMPA (Towards Richer Online Music Public-domain Archives) H2020 project.
- Paradigms of Artificial General Intelligence and Their Associated Risks, project at Centre for the Study of Existential Risks, University of Cambridge.

References

1. Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner, 'Machine bias', ProPublica, 2016.
2. Adrien Bennetot, Vicky Charisi, and Natalia Díaz-Rodríguez, 'Should artificial agents ask for help in human-robot collaborative problem-solving?', arXiv preprint arXiv:2006.00882, 2020.
3. Vicky Charisi, Alyssa M. Alcorn, James Kennedy, Wafa Johal, Paul Baxter, and Chronis Kynigos, 'The near future of children's robotics', in Proceedings of the 17th ACM Conference on Interaction Design and Children, IDC '18, p. 720–727, New York, NY, USA, 2018.
4. Vicky Charisi, Emilia Gómez, Gonzalo Mier, Luis Merino, and Randy Gomez, 'Child-robot collaborative problem-solving and the importance of child's voluntary interaction: A developmental perspective', Frontiers in Robotics and AI, 7, 15, 2020.
5. Vicky Charisi and Kristiina Jokinen, 'Expressive multimodal communication for human-robot interactions in cross-cultural settings', in ACM/IEEE International Conference of Human-Robot Interaction, 2019.
6. Vicky Charisi, Laura Malinverni, Elisa Rubegni, and Marie-Monique Schaper, 'Creating opportunities for children's reflections on AI, robotics and other intelligent technologies', in Proceedings of the 19th ACM Conference on Interaction Design and Children, IDC '20, p. 720–727, New York, NY, USA, 2020.
7. Vicky Charisi, Selma Sabanovic, Serge Thill, Emilia Gomez, Keisuke Nakamura, and Randy Gomez, 'Expressivity for sustained human-robot interaction', in 2019 14th ACM/IEEE International Conference on Human-Robot Interaction, pp. 675–676, 2019.
8. Shai Danziger, Jonathan Levav, and Liora Avnaim-Pesso. 'Extraneous factors in judicial decisions' Proceedings of the National Academy of Sciences 108, no. 17, 2011.
9. Enrique Fernández-Macías, Emilia Gómez, José Hernández-Orallo, Bao Sheng Loe, Bertin Martens, Fernando Martinez-Plumed, and Songül Tolan, 'A multidisciplinary task-based perspective for evaluating the impact of AI autonomy and generality on the future of work', arXiv preprint arXiv:1807.02416, 2018.
10. Ana Freire, Lorenzo Porcaro, and Emilia Gómez. 'Measuring diversity of artificial intelligence conferences', arxiv preprint arXoov:2001.07038, 2020.
11. Emilio Gómez-González and Emilia Gómez, 'Artificial intelligence in medicine and healthcare: applications, availability and societal impact', Technical report, EUR 30197 EN, Publications Office of the European Union, Luxembourg, 2020, ISBN 978-92-76-18454-6 (online), doi:10.2760/047666 (on- line), JRC120214, 2020.
12. José Hernández-Orallo, 'Evaluation in Artificial Intelligence: from task-oriented to ability-oriented measurement', Artificial Intelligence Review, 48(3), 397–447, 2017.
13. Daniel Kahneman, Thinking, fast and slow, Macmillan, 2011.
14. Martínez-Plumed, Fernando, Gómez, Emilia, and José Hernández-Orallo, 'Tracking the evolution of AI: The AI-collaboratory', in Proceedings of the 1st International Workshop: Evaluating Progress in Artificial Intelligence (EPAI), 2020.
15. Fernando Martínez-Plumed and José Hernández-Orallo, 'Dual indicators to analyse AI benchmarks: Difficulty, discrimination, ability and generality', IEEE Transactions on Games, 12 (2), pp 121 - 131, 2020.
16. Fernando Martínez-Plumed, Shahar Avin, Miles Brundage, Allan Dafoe, Sean Ó hÉigeartaigh, and José Hernández--Orallo, 'Accounting for the neglected dimensions of AI progress', arXiv preprint arXiv:1806.00610, 2018.
17. Fernando Martínez-Plumed, Emilia Gómez, and José Hernández-Orallo, 'Tracking AI: The capability is (not) near', in Frontiers in Artificial Intelligence and Applications. Volume 325: ECAI 2020. IOS Press, pp 2915 - 2916, 2020.
18. Fernando Martínez-Plumed, Songül Tolan, Annarosa Pesole, José Hernández-Orallo, Enrique Fernández-Macías, and Emilia Gómez, 'Does AI qualify for the job? A bidirectional model mapping labour and AIintensities', in Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, AIES '20, p. 94–100, New York, NY, USA, 2020.
19. Marius Miron, Songül Tolan, Emilia Gómez, and Carlos Castillo, 'Evaluating causes of algorithmic bias in juvenile criminal recidivism', Artificial Intelligence and Law, 1–37, 2020.
20. Lorenzo Porcaro, Carlos Castillo, and Emilia Gómez, 'Music recommendation diversity: A tentative framework and preliminary results', in 1st Workshop on Designing Human-Centric MIR Systems, colocated at 20th ISMIR Conference, 2019.
21. Lorenzo Porcaro and Emilia Gómez, '20 years of playlists: A statistical analysis on popularity and diversity', in 20th Conference of the International Society for Music Information Retrieval (ISMIR 2019), Delft, The Netherlands, 2019.
22. Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin, 'Model-agnostic interpretability of machine learning', arXiv preprint arXiv:1606.05386, 2016.
23. Sergio Sánchez-Martínez, Oscar Camara, Gemma Piella, M. Cikes, M.A. Gonza´lez Ballester, Marius Miron, A. Vellido, Emilia Gomez, A. Fraser, and Bart Bijnens, 'Machine learning for clinical decision making Challenges and opportunities', Preprints 2019110278.
24. Bob L. T. Sturm, Maria Iglesias, Oded Ben-Tal, Marius Miron, and Emilia Gómez, 'Artificial intelligence and music: Open questions of copyright law and engineering praxis', Arts, 8(3), 115, 2019.

25. Songül Tolan. „Fair and unbiased algorithmic decision making: Current state and future challenges.“ arXiv preprint arXiv:1901.04730, 2019.
26. Songül Tolan, Marius Miron, Emilia Gómez, and Carlos Castillo, ‘Why machine learning may lead to unfairness: Evidence from risk assessment for juvenile justice in Catalonia’, in Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law, ICAIL ’19, pp. 83–92, New York (USA), ACM, 2019.
27. Songül Tolan, Annarosa Pesole, Fernando Martínez-Plumed, Enrique Fernández-Macías, Emilia Gómez, et al., ‘Measuring the occupational impact of AI: Tasks, cognitive abilities and ai benchmarks’, Technical report, Joint Research Centre, 2020.
28. Daniel Ullman and Bertram F Malle, ‘Measuring gains and losses in human-robot trust: evidence for differentiable components of trust’, in 2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI), pp. 618–619. IEEE, 2019.
29. Francisco J Varela, Evan Thompson, and Eleanor Rosch, The embodied mind: Cognitive science and human experience, MIT press, 2016.
30. Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork, ‘Learning fair representations’, in International Conference onMachine Learning, pp. 325–333, 2013. http://proceedings.mlr.press/v28/zemel13.html
31. Martínez Plumed, F.; Tolan, S.; Pesole, A.; Hernández Orallo, J.; Fernández-Macías, E.; Gómez, E. ‘Does AI Qualify for the Job?A Bidirectional Model Mapping Labour and AI Intensities (Appendix)’. http://hdl.handle.net/10251/133314, 2020.
32. Martínez-Plumed, F., Ferri, C., Nieves, D., and Hernández-Orallo, J. ‘Fairness and Missing Values’. arXiv preprint arXiv:1905.12728, https://arxiv.org/abs/1905.12728 2019
33. De Nigris, S., Gomez-Gonzales, E., Gomez Gutierrez, E., Martens, B., Iglesias Portela, M., Vespe, M., Schade, S., Micheli, M., Kotsev, A., Mitton, I., Vesnic Alujevic, L., Pignatelli, F., Hradec, J., Nativi, S., Sanchez Martin, J.I., Hamon, R. and Junklewitz, H., ‘Artificial Intelligence and Digital Transformation: early lessons from the COVID-19 crisis’, Craglia, M. editor(s), EUR 30306 EN, Publications Office of the European Union, Luxembourg, 2020, ISBN 978-92-76-20802-0 (online), doi:10.2760/166278 (online), JRC121305.
34. Eren, Ozkan, and Naci Mocan. 2018. ‘Emotional Judges and Unlucky Juveniles’. American Economic Journal: Applied Economics, 10 (3): 171-205.

JRC122667

European Commission, 2020

How to cite this report: Gomez Gutierrez, E., Charisi, V., Tolan, S., Miron, M., Martinez Plumed, F. and Escobar Planas, M., Centre for Advanced Studies, Amran, G. editor(s), Publications Office of the European Union, Luxembourg, ISBN 978-92-76-28212-9, doi:10.2760/23970, JRC122667.

The authors would like to thank Stephan Lindner for his work on layout and design of the printed version.

GETTING IN TOUCH WITH THE EU

In person

All over the European Union there are hundreds of Europe Direct information centres. You can find the address of the centre nearest you at: https://europa.eu/european-union/contact_en

On the phone or by email

Europe Direct is a service that answers your questions about the European Union. You can contact this service:

- by freephone: 00 800 6 7 8 9 10 11 (certain operators may charge for these calls),
- at the following standard number: +32 22999696, or
- by electronic mail via: https://europa.eu/european-union/contact_en

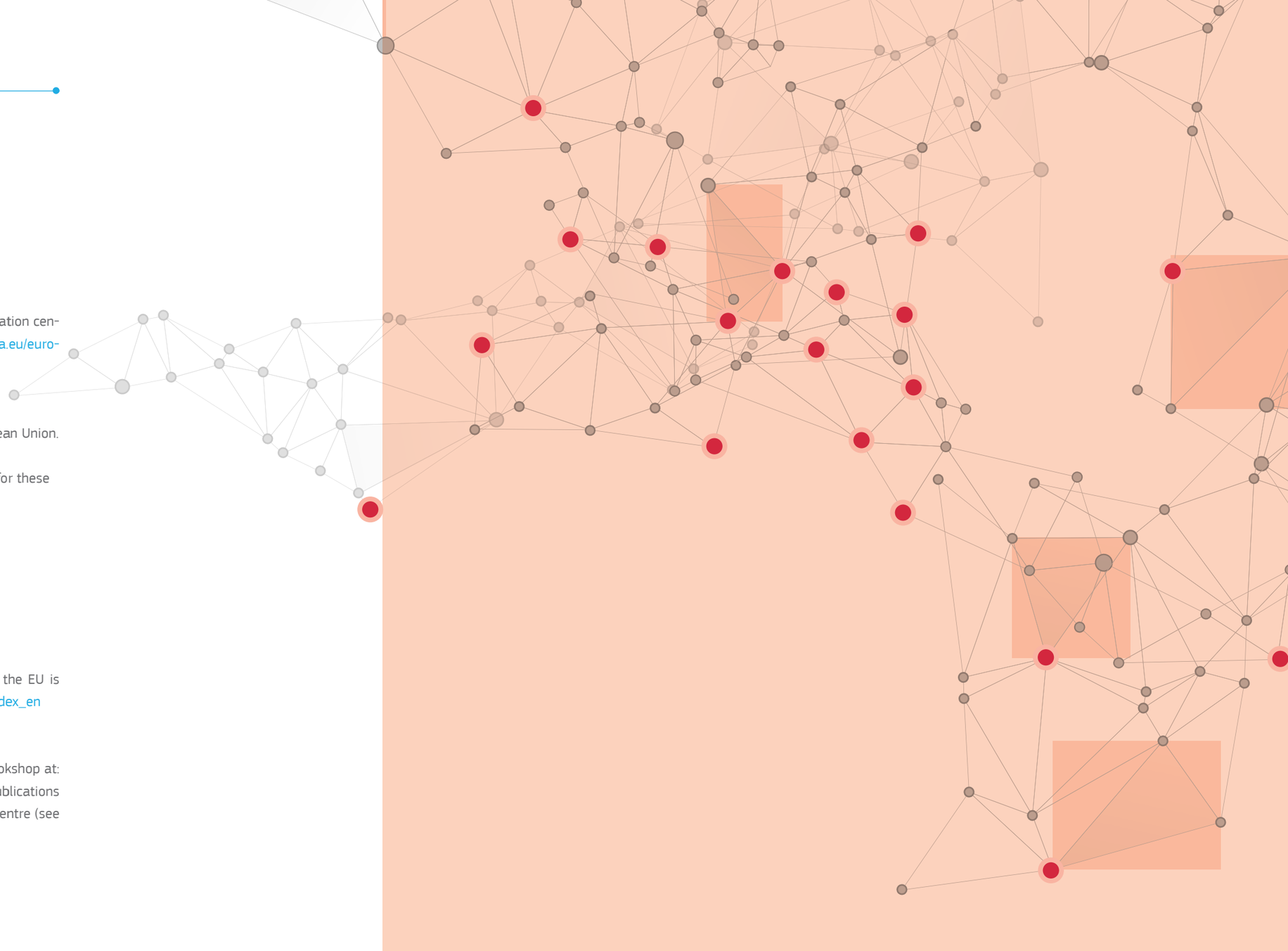
FINDING INFORMATION ABOUT THE EU

Online

Information about the European Union in all the official languages of the EU is available on the Europa website at: https://europa.eu/european-union/index_en

EU publications

You can download or order free and priced EU publications from EU Bookshop at: https://publications.europa.eu/en/publications. Multiple copies of free publications may be obtained by contacting Europe Direct or your local information centre (see https://europa.eu/european-union/contact_en).



KJ-04-20-745-EN-N

The European Commission's science and knowledge service

Joint Research Centre

JRC Mission

As the science and knowledge service of the European Commission, the Joint Research Centre's mission is to support EU policies with independent evidence throughout the whole policy cycle.

EU Science Hub

ec.europa.eu/jrc

@EU_ScienceHub

EU Science Hub- Joint Research Centre

EU Science, Research and Innovation

EU Science Hub

Publications Office of the European Union

doi:10.2760/23970
ISBN 978-92-76-28212-9