USING ABM AND SERIOUS GAMES TO CREATE "BETTER AI"

Petra Ahrweiler[a], Nigel Gilbert[b], Zsolt Juranyi[c], Martha Bicket[b],
Albert Sabater Coll[d], George Kampis[e], Blanca Luque Capellas[a], and David Wurster[a]

[a]TISSS Lab, Johannes Gutenberg University Mainz, Germany
{*petra.ahrweiler, bluqueca, dwurster*}*@uni-mainz.de*
[b]CRESS, University of Surrey, United Kingdom
{*n.gilbert, m.bicket*}*@surrey.ac.uk*
[c]DeepData Kft., Hungary
juranyi@deepdata.hu
[d]Faculty of Business and Economic Sciences, University of Girona, Spain
albert.sabater@udg.edu
[e]DFKI German Research Center for Artificial Intelligence, Germany
george.kampis@dfki.de

ABSTRACT

The paper focuses on the challenge of providing contextualized, value-sensitive, and participatory Artificial Intelligence (AI) that meets societal needs. The "AI for social Assessment" (AI FORA) project combines empirical research, gamification, and agent-based models (ABM) to assess the fairness of AI-based distribution in different countries and propose improvements. The paper presents a case study, where ABM and serious games are used to identify more desirable social assessment routines. A machine learning (ML) system is developed. The workflow for prototyping AI social assessment systems involves creating a synthetic population, generating synthetic outcomes based on improved assessment rules, training a neural network, and evaluating the effectiveness of the AI system. By following this approach, the paper suggests that unintended consequences and ineffective systems can be avoided, saving on costly development. The aim is to ensure that AI for social service delivery is responsive, fair, and beneficial to society.

Keywords: social assessment, public service provision, ML and training data, context-specific ABM

1 INTRODUCTION

Public administrations are increasingly using Artificial Intelligence (AI) algorithms to decide on the provision of public social services such as unemployment benefits, pension entitlements, kindergarten places and social assistance to their citizens, hoping to achieve greater efficiency and objectivity [1, 2]. Data profiles of citizens are analyzed and assessed, and profiles automatically checked and scored to determine whether their owners are eligible to receive support from the state. However, AI-based social assessment systems, because they are based on machine learning (ML) from historical data, are accused of perpetuating bias and discrimination, often to the detriment of the most vulnerable groups in society. Who is considered as eligible, needy and deserving to be a beneficiary will always imply decisions that privilege certain groups while discriminating against others. Criteria vary widely around the world. There is no approach to social assessment that would be perceived as fair everywhere. Fairness concepts vary across national welfare systems depending on culture, religious tradition, and belief system [3, 4, 5].

Proc. of the 2024 Annual Simulation Conference (ANNSIM'24), May 20-23, 2024, Washington D.C., USA
C. Ruiz-Martin, B. Oakes and R. Cárdenas, eds.

How can contextualized, value-sensitive, responsive and dynamic AI systems be co-designed starting from existing systems that are perceived as problematic? The 'AI for Assessment' (AI FORA) project combines empirical research on AI-based social service delivery with gamification at community-based multi-stakeholder workshops and a series of case-specific agent-based models for assessing the status quo of AI-based distribution fairness in different countries, for simulating desired policy scenarios, and for generating an approach to 'Better AI'.

Thus, there is, so far, no way to prototype better AI for advising social workers. Important questions of great policy relevance cannot be put to the test: What would happen if the stakeholders' insights into the flaws and gaps of current assessment practices were used to revise the rules of decision making? How would the population distribution of social services change? Which combination of rules would allow for the best outcomes?

The paper is structured as follows: In Section 2 we give a broad overview of the overall AI FORA modelling approach and introduce the AI FORA Spanish case study, which will be used in the remainder of the paper to illustrate the key components of this approach in practice. Section 3 introduces the case-specific game design and ABM approach to identify more desired ways of social assessment routines [6]. Section 4 specifies the need for a ML system advising social workers on options for assessing applicants. Section 5 describes our workflow for realizing such a system and how it could be used for prototyping AI social assessment systems. Our conclusions and next steps are outlined in Section 6.

2 BACKGROUND TO THE AI FORA APPROACH AND SPANISH CASE STUDY

2.1 The AI FORA modelling approach

A participatory modelling strategy was designed to support the transition from existing to desired social assessment systems, with the following elements for each case study:

1. A workshop is held to map out the overall existing case study system as a flow chart.
2. An Agent-Based Model (ABM) that models the current social assessment system, including an initial ruleset (ruleset version 1) and exemplar agent attributes, is written.
3. Ruleset version 1 is checked and iteratively refined by running the ABM.
4. Rules based on the ABM for a game to be played with stakeholders are written.
5. At a gamification workshop with the stakeholders, the rules are gradually adapted by the stakeholders.
6. A 'better ruleset', ruleset version 2, is extracted using the records from the game play.
7. A synthetic population is created to match the real population on relevant attributes.
8. Ruleset version 2 and the synthetic population are used to generate a training dataset.
9. A machine learning system is trained using the training data which could be used to assess applicants.

Figure 1 depicts the modelling process, elements of which are now introduced in more detail.

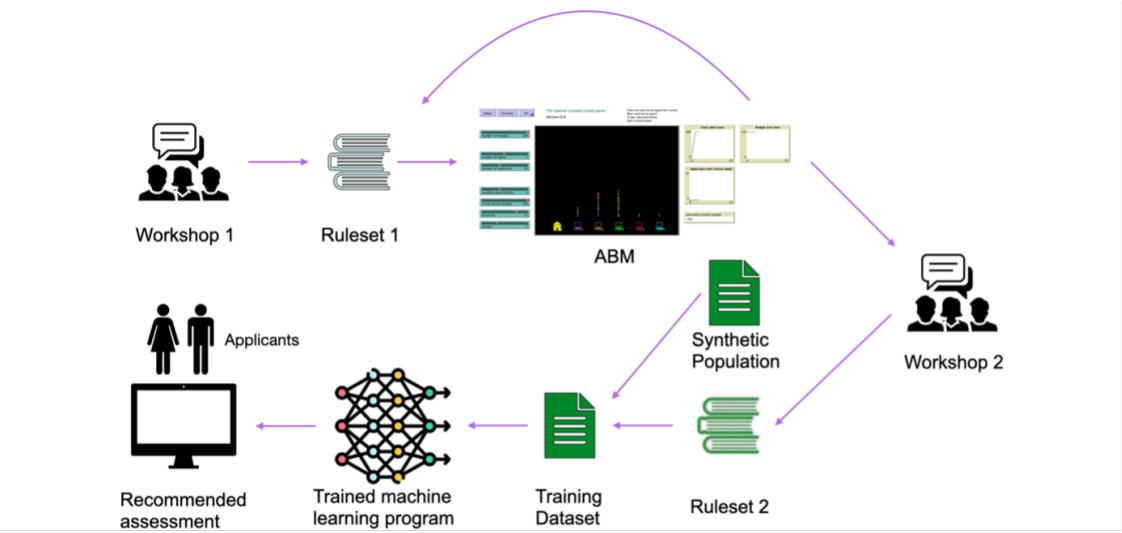

Figure 1: Illustration of the modelling strategy.

2.1.1 Review of existing systems

A detailed system map about how social assessment routines for distributing social services are conceptualized, organized and institutionalized in the case study country is created. This includes a policy analysis and a technical analysis for each context. Mapping the existing actor network requires research, both quantitative and qualitative, complemented by Participatory Systems Mapping [7] to reconstruct the existing system from the perspective of stakeholders. This work provides information on the actors involved, the societal norms and values, the organizational practices and routines in place, and the system's performance. This data is used to create an ABM representation of the existing social assessment routines, which both informs, and is iteratively improved upon by, the design of a game to be played with stakeholders.

2.1.2 Gamification as a method

Interactive and participatory formats at multi-stakeholder workshops expose the culturally shaped and heterogeneous value perspectives of the local social groups. As a central component, participants play 'serious games' for co-designing better AI systems [8, **?**, 9]. Gamification, i.e., applying game elements in non-game contexts [10], is a low-threshold entry point for non-scientists to contribute to research. The games mimic a social assessment situation in distributing public services. They are designed to explore how people from different backgrounds would create social assessment algorithms that are better from their cultural perspective. Games create a controlled setting with observability, measurability and comparability. They complement data collection on desired scenarios for better AI systems from a stakeholder perspective [11, 12] and help to identify questions for scenario simulations. Stakeholders suggest, discuss, co-develop and test interventions in all parts of the game situation, including the social assessment criteria ('changing the algorithm') to propose a 'better' ruleset for that assessment system. By 'ruleset' we mean a collection of rules that when followed (by a clerk or by agents in the ABM) can be used to classify an applicant as deserving of full, partial or no social services allocation. These rules will take account of the applicant's situation and attributes. The gamification approach empowers stakeholders to tackle the problem of distributing scarce resources in the context of their specific socio-cultural setting. The advantage of iterating between games and ABM is that stakeholders can deliberate in the game context with options to 'change the rules of the game' as result of their discussions, the ABM then codifies and formalizes the rules showing the results of their application in the game environment, which, in the next iteration, informs stakeholder discussions in the game.

2.1.3 Anticipating, projecting and realizing desired systems

Once the existing and 'better' ruleset have been determined through the iterative process of building ABMs and running gamification workshops for each case study country, these rulesets can be used for the next step in the modelling process. Although the outcomes of the rulesets have been already demonstrated by the ABM, this is in only a small game environment. It does not show what the rulesets would do to the whole population of case study countries, especially not given population dynamics. This is why we now leave the world of games and ABM and enter the world of population databases, neural networks and ML.

Rather than using micro-data for the population itself, to comply with research ethics and data protection issues, a synthetic population database is generated that resembles the real-world data for a case study country. First, the ruleset of the existing system is run on the synthetic population. It produces an output database with some individuals getting service and others not. To validate this, we could check against real-world data on the socio-demographics of social service recipients; however, in most cases such data will not be available. Instead, we can check whether the database reproduces the number of service recipients in the real world (if available), whether it reproduces the stylized facts on bias and discrimination in the literature, and whether it reproduces the case study's empirical research results. Second, the ruleset of the desired system is run on the synthetic population. It produces an output database with different distributions from the first. Stakeholders (and decisionmakers) can now compare these two "data worlds" and check the effects of different assessment algorithms on the overall population and the expected population dynamics. Once the assessment algorithm is confirmed, the corresponding database can be used as training data for a neural network that recommends social assessment decisions for distributing services in the real-world context.

2.2 The Spanish case study

The main goal of the AI FORA Spanish case study is to examine the perceptions, attitudes and acceptance of AI-based social assessment technologies by policy makers and administrative agencies locally in Catalonia, a frontrunner Spanish region in the adoption of digital technologies for the public sector. A mixed-methods methodology which involves gamification workshops, focus groups, and discourse analysis with in-depth interviews with a variety of stakeholders and media sources from state, regional and local news, was employed. Given the overrepresentation of vulnerable groups who use social services in the cities of Barcelona, Girona and Mataró, special emphasis was given to the impact of AI systems on such groups, particularly migrant populations. Since poverty, social exclusion and vulnerability require information and access to various data sources to make a proper assessment of an individual's social status, a digital assessment tool supports social service clerks in the diagnosis and detection of complex cases and guides the intervention and follow up on each individual case. The tool is called the 'Self-Sufficiency Matrix Catalunya (SSM-Cat)' by municipalities in Catalonia. It originated from The Netherlands [13] and the USA [14] and was adapted and validated by the Department of Work, Social Affairs and Families of the Generalitat de Catalunya, in collaboration with Municipal Associations and the College of Social Work and the College of Educators and Social Educators. It is now in use in all municipalities with the aim of unifying an assessment system that, until recently, only existed in some localities.

While previous methods of assessment for social provision of complex needs were subjective, depending on the opinions of social service clerks, the SSM-Cat reduces discretion by measuring an individual's degree of self-sufficiency along 13 dimensions. In doing so, the tool reduces the social worker's task to obtaining a relatively simple view of complex social needs. The introduction of the SSM-Cat at the local level had two main objectives: (1) to increase transparency in the decision-making process; and (2) to provide a more consistent and comparable tool for monitoring the allocation of social services across municipalities whilst striving to enhance fairness in social service provision.

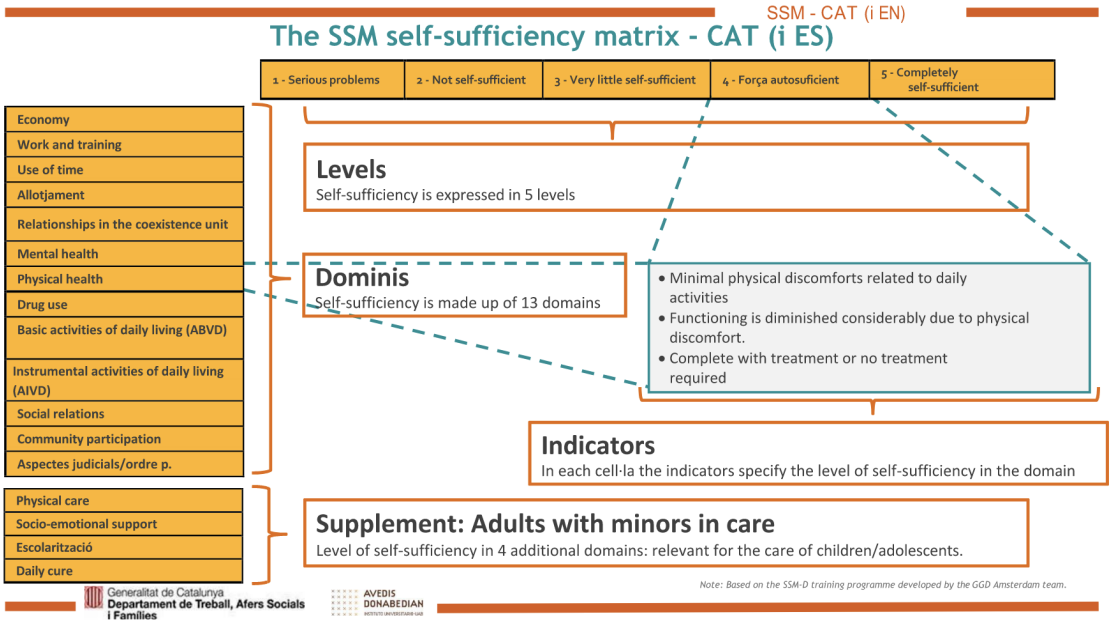

Figure 2: The Self-Sufficiency Matrix (SSM).

3 ABM AND SERIOUS GAMES FOR IMPROVING SOCIAL ASSESSMENT PRACTICES

In this section we illustrate the participatory modelling strategy outlined above by describing how it has been applied to the Spanish case study.

3.1 The Spanish ABM

An agent-based model (ABM) was developed to represent the Catalonian SSM-Cat social assessment system. Here we describe the main features of the ABM using the framework proposed in [15]. The main premise of the ABM is that social service clerks are working in a municipality. They work at desks and seek to allocate limited social service resources to deserving applicants, many of whom have multiple, complex needs. The clerks' aim is to identify and allocate social service resources to applicants to maximize the wellbeing of applicants.

3.1.1 Agents

The ABM agents are the applicants hoping to be seen and allocated budget by the clerks. For the sake of simplicity, the ABM reduces the SSM-Cat's 13 dimensions of wellbeing to 7 dimensions in total. Six of the applicants' attributes can vary over time during the ABM simulation: household income, accommodation, work and training, mental health, physical health, and an aggregate 'overall need score', described in more detail below. A seventh attribute, number of dependents, is fixed from initialization.

3.1.2 Environment

Applicants can be at home (at the end of each round), or in a queue in front of a clerk's desk. The following global attributes define the ABM environment in which the clerks and applicants interact:

- Number of applicants
- Number of clerks
- Number of the round, starting with zero
- Social services budget measured in money units, which is refreshed at the end of each round
- Available appointments (the number of applicants that clerks can see each round)
- Threshold (the threshold above which applicants are considered critically-in-need). If any applicant's need score is above this threshold, the next round's budget is reduced.

The clerks can observe the budget, the number of applicants queuing at their desk and the threshold.

3.1.3 Actions and interactions

The clerks carry out social assessments of applicants at their desks. They review and score applicants based on an algorithm that depends on the ruleset and applicants' attributes.

Under the initial ('existing') ruleset, the algorithm is as follows: For the attributes 'household income' and 'number of dependents', applicants are ranked against each other and given between 1 and 5 need points based on their position relative to other applicants that round. For the remaining attributes, the points assigned to applicants by clerks are shown in Table 1. An applicant's 'overall need score' is then calculated to be the sum of points for each attribute.

Table 1: Need points for scoring.

Need points	5	4	3	2	1
Applicants score	Serious problems	Not self-sufficient	Minimally self-sufficient	Sufficiently self-sufficient	Completely self-sufficient

Receiving support alleviates an applicant's worst need-category (decreasing it by 1). If an applicant did not receive any support in that round, and if they had any categories where they had a score greater than or equal to 4 (not self-sufficient) then those and one other attribute, chosen at random, gets worse (their scores are increased by 1). An applicant may then experience a random change in fortune before the next round. There is a 10% chance that one of their attributes will worsen by 1 and a 10% chance that one of them will improve by 1. When critically needy applicants' requirements are not met in a round, this impacts the upcoming round's available budget because these applicants will draw on social services elsewhere in the system. However, the applicants' need scores are not improved.

3.1.4 Temporality

These interactions repeat for a set number of rounds. If an applicant's score is good enough (<= 2), applicants stay at home for a round rather than visiting a clerk. Applicants in need of an appointment are evenly distributed to clerks at the beginning of each round. At the end of the round the social service budget is distributed to successful applicants in order of severity: the highest scoring applicant is paid, then second highest applicant and so on until the budget for that round is used up. At the end of each round, applicants' needs are updated depending on their existing needs and whether they received support. The run ends if there is no budget left at the beginning of a round to allocate to applicants.

3.1.5 NetLogo implementation and interface

The model is implemented in NetLogo and available at https://github.com/micrology/AIFORA/tree/main/Games/Spain. Figure 3 depicts the interface after ten rounds.

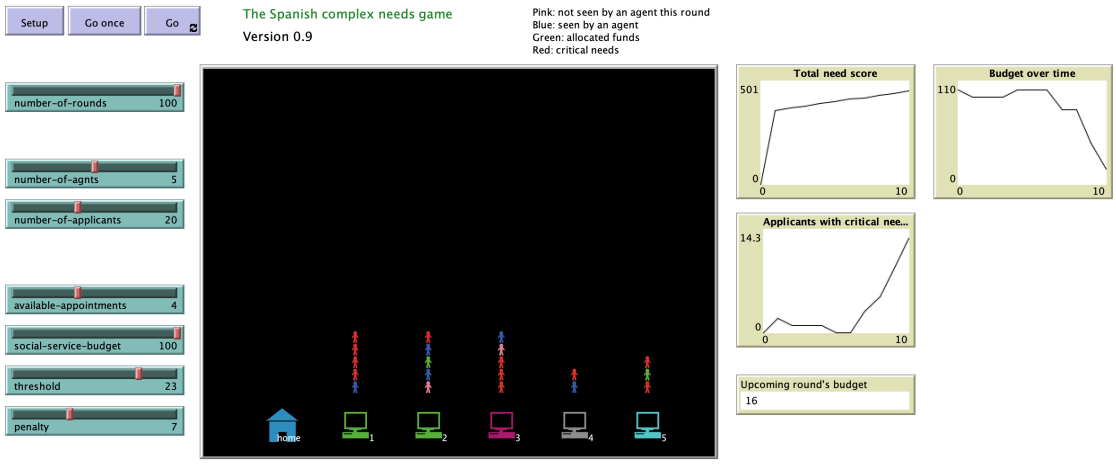

Figure 3: Spanish case study ABM.

A home station and five clerk desks can be seen, with twenty applicants distributed between the desks. The maximum number of rounds, number of desks, applicants, available appointments per clerk (currently set at 4), available budget, threshold for critical needs, and the penalty for having too many applicants with critical needs can all be changed with sliders. Blue applicants have been seen by the clerk that round but have not received support; green applicants have received support; and red applicants have critical needs. Pink applicants were not seen by a clerk in the current round. As can be seen from the plots on the right, the system is not in good shape: the budget is nearly gone, there is a high total need score, and there are many applicants with critical needs.

3.2 The Spanish game

The role of the ABM described in the previous section is to act as a kind of theorem-checking device for the ruleset derived from the one in place in the empirical system under investigation, in this case social assessment in Catalonia. The ABM ensures that the ruleset is coherent and complete, and it acts as an informed starting point for devising a better algorithm. However, the ruleset does not capture informal practice, might be considered as unfair, and might not produce desired system features. Improving the ruleset requires stakeholder involvement. In Catalonian municipalities, the assessment of individuals seeking social assistance is not further standardized beyond the SSM-Cat. An individual's assessment depends entirely on the social workers' perceptions and experiences in interpreting and applying the SSM-Cat. Therefore, a gamification workshop was designed and held at Montserrat Abbey in May 2023 to investigate assessment behavior using the SSM-Cat in practice and to explore possible improvements to the system. Workshop participants consisted of public managers responsible for policy implementation of a full range of design tools and methods to assess social services, and social service workers from Barcelona, Girona and Mataró as well as practitioners in social service provision from local NGOs such as Caritas and the Red Cross, all of whom had a good understanding of the Self-Sufficiency Matrix and experience in using it.

3.2.1 Purpose of the game

Prior research had shown that stakeholders' ideas for improvements to the system varied greatly. The workshop game sought to investigate the variety of criteria used in decision making for assessment with the SSM-Cat, to examine whether social workers were able to develop an 'interpretation culture' about fairness issues, especially concerning vulnerable populations, and to see whether they would converge in judging applicants' profiles. Although social workers sometimes struggled for consistency in their interpretation of

the high-needs profiles, the game demonstrated that it is generally possible to identify specific guidelines or rules related to an 'interpretation culture' that help them make more accurate and consistent judgments across different applicants or profiles and situations.

3.2.2 Game design

Workshop attendees play ten social service clerks working in a municipality. They sit in pairs at five service desks, one playing the responsible officer, the other an office helper supporting procedures at the desk, thus allowing his/her colleague to concentrate on an applicant. The applicants, twenty individuals requesting social service support, are played by researchers present at the workshop. Each applicant has an identity card with a short biographical narrative telling their story and explaining their profile.

Home: Clerks and applicants start and end each round of the game (a 'day') at home. At the start of each day, applicants receive their identity cards and familiarize themselves with its narrative. In each round, identical profiles are presented at all the desks. Clerks are randomly assigned to desks with office helpers who support filling in the self-sufficiency matrix. Applicants are distributed evenly between the desks. At the end of the day, clerks and applicants return home and complete diary entries about their day and their situation.

Desks: The clerks administer the state budget, which decreases after each round of the game according to the services provided. Clerks evaluate applicants who arrive at their desk and tell their story based on their narrative. Clerks assess these applicants by translating their narrative into scores for each of the categories of the self-sufficiency matrix and allocate either full social service provision, partial social service provision, or no social service provision, depending on their assessment. The helpers support the officers to complete the assessment sheets with the score for each applicant. After seeing a clerk, applicants go home. Once they have assessed all of the applicants at their desk, the clerks go to the Office Meeting Place. Clerks can run out of budget and then must send applicants home empty-handed.

Office Meeting Place: Clerks interact at the Office Meeting Place, moderated by a local office manager. They bring their assessment notes and can deliberate about the features of the Self-Sufficiency Matrix and their individual assessments of applicants. Clerks propose new guidelines for how profiles should be assessed. A vote about whether those guidelines will be implemented is taken at the end of the round before clerks leave for home. If the proposed guidelines are accepted, the Self-Sufficiency Matrix is amended accordingly for the next round and distributed to all service desks.

The game continues as long as there is at least one desk with budget. When the budget has run out completely, the game ends.

3.2.3 Conclusions drawn from playing the game

Stakeholder deliberations revealed that there were certain criteria that all social workers use in assessment but which are missing from the SSM-Cat. They discussed whether these criteria should be added to the SSM-Cat to align with current practice, or whether this current practice should be discouraged. Furthermore, discussions confirmed that all stakeholders were aware of past bias and discrimination issues in their country and actively tried to avoid them. They were surprised, however, to discover the degree to which their assessments of applicants varied, even though they were all presented with the same profile, indicating that the SSM-Cat was interpreted in different ways by the social workers. Stakeholders expressed their interest in developing a common interpretation culture in agencies about how to apply the SSM-Cat, and that this could be supported by using profiles and narratives in training, as was done in the game. In all, stakeholders engaged in collaborative decision-making processes through discussions, shared insights, and worked together to establish a shared understanding of the rules and criteria for evaluating high-needs profiles.

4 AI FOR ADVISING SOCIAL WORKERS ON OPTIONS FOR ASSESSING APPLICANTS

In all our case study countries, as in Spain, the existing social assessment systems, whether based on AI or merely involving the digitalization of previously manual systems, were seen to be biased and to discriminate against certain groups, especially vulnerable people. This is confirmed by the literature on the individual national welfare systems and the evaluation of databases on social service decision making [16]. Since AI is increasingly being proposed as a way of making social assessment more efficient, quicker and less costly, it is important to devise ways of making such systems fairer, while recognizing that 'fairer' is a culturally specific notion. The challenge that the AI FORA project is aiming to contribute to is how to design social assessment technology that is value-sensitive, contextualized, dynamic and responsive to social change. Legislation in most countries disallows automated decision making based on scoring (e.g. [17, 18, 19]). However, according to AI FORA's case studies, AI could be used in social assessment if it were to advise social workers on options for assessing applicants.

The work that we have done, using an interacting cycle of agent-based modelling and serious games, points to an approach by which the technology can be specified in a stakeholder-driven way, so that it is more transparent and discursive about bias and discrimination, includes values such as the social justice concept of the society in which it will be used, and is responsive to the needs of vulnerable groups. How, exactly, can this be realized? It is well documented that AI technology may not only provide biased information, but may also inadvertently reinforce existing cultural, social, and economic inequalities. Thus, if an AI system is trained on data that reflects unequal access to resources or opportunities, it may further entrench these disparities by providing advantages to dominant groups. The lack of social awareness of the assessment provided by predictive tools is particularly problematic because of the nature of decisions that can be based on them, such as the deployment of social services and targeting problem-oriented responses to the right persons and places.

This problem is especially important when used in the context of social assessment. "It's the data, stupid!" is an often-cited phrase indicating that, though data is the problem, data is data and cannot be changed. The algorithm would be able to learn for less bias and discrimination, but the world is as it is: Bias and discrimination is an intrinsic feature of decision making producing the data that trains the algorithm. We can dream of a less biased and fairer world, but unfortunately it is not available. Therefore, we will not get better data to train a better algorithm as "better AI" – even if decision makers, e.g. social workers deciding on concrete individual cases, would be willing to avoid bias and discrimination that have been identified and discussed as undesired for the future of their systems. The problem is reinforced by the fact that the few, mostly US-based, companies providing software in this area work with datasets from their own countries: They export de-contextualized technology trained on data reflecting bias and discrimination in their home country without any reference to the cultural values of the technology-using countries.

In contrast, our approach of combining games and ABM for stakeholder input on desired systems offers the possibility of producing relevant training data that can be used to prototype algorithms fitting desired futures.

5 A WORKFLOW TO CREATE 'BETTER AI'

This section offers a method for prototyping AI social assessment systems by illustrating how rulesets extracted from the ABM and gamification workshops in section 3 could be used within an AI-based social assessment system. Continuing with the example of the AI FORA Spanish case study, we outline our approach to training a neural network on both the existing and stakeholder-revised rulesets, using a synthetic population.

5.1 Creating a synthetic population database

To comply with research ethics and data protection issues, a synthetic population database is generated that resembles the real-world data for a case study country.

To create the synthetic population, we draw on sociodemographic data about the Catalan population and relate it to the agents' attributes mentioned in section 3.1. Data was collected in 2020, except data related to accommodation, collected in 2014. The following sources are used: the Institut d'Estadística de Catalunya [20], the report "Quantificació i distribució territorial de la població mal allotjada a Catalunya: Informe de resultats" [21] and the report "Informe sobre l'estat dels serveis socials a Catalunya 2020" [22].

From these sources, we draw sociodemographic data about the Catalonian population for the variables: "Population structure by sex and age groups" and "Foreign population by age and sex". Data about agents' attributes (as described in section 3.1) is derived from the following variables: for "Household income": "Composition of the at-risk-of-poverty population by sex and age"; for "Accommodation": "Cases of bad accommodation dealt with in Catalonia"; for "Work and training": "Unemployed population", "Wage-earning employed population. Type of contract" and "Population aged 15 and over by level of education attained"; for Mental health" and "Physical health": "Persons legally recognised as disabled according to degree of disability", "Persons legally recognised as disabled according to type of disability", "Persons legally recognised as disabled by gender" and "Persons legally recognised as disabled on the basis of age"; and for "Number of dependents": "Homes with one family or more, by type of family and number of children under 16 years of age".

Taking the original population database, we generate a synthetic population of the same size with the same statistical distribution. For the Spanish case study, N equals 7,732,756, the total Catalan population in 2020. To do this, a table with N empty rows is generated, then columns of the variables are added. The independent attributes (e.g. age) are added first as there are others depending on them. For each category of a variable, K randomly selected synthetic individuals receive that category, where K is the known count of real individuals in that category. For dependent variables (e.g. employment depends on age), the same procedure is applied with the difference that the random rows are picked from a filtered set: the relevant rows (e.g. with age 16-64 for employment), instead of the whole synthetic population. This way, we can produce a synthetic population of any size where issues of privacy do not arise.

After following the process above for the Spanish case study, the resulting table now contains data on 7,732,756 synthetic individuals whose attributes mimic the real-world statistics. This dataset can be used to simulate the social assessment for public service provision.

It can be beneficial to simulate the aging of the synthetic population as well, to predict possible future states. Doing the simulations and applying the workflow described in this paper can aid stakeholders work ahead of time on adjusting rulesets in order to better serve the future state of society.

5.2 The ruleset for social assessment

For scoring the individuals in the synthetic database, we go back to the ABM-generated rulesets on social assessment. We want to check how these rulesets perform at a population level, and not only in a game environment as in the ABM.

First, the ruleset of the existing system, generated by the iterative process between games and ABM, is run on the synthetic population.

As noted in section 3, social assessments are typically made by considering applicants' characteristics and histories, for example, in the Spanish case, by considering their income, accommodation, labor market ex-

perience, and health. Social workers have to make judgements about an individual's scores on each of these features in preparation for making an assessment. Inevitably, such judgements are to a degree subjective and prone to some degree of error or bias. We reproduce this uncertainty in processing the simulated population. As explained above, the synthetic population includes, for each agent in the population, a set of attributes that were used in the game and that are relevant to making an assessment (i.e., household income, accommodation, work and training, and physical and mental health). Most of these attributes are categorical, i.e. have one of a small number of possible values, or use 5-point Likert scales, but the training data needs to have numerical values. We therefore translate the categories to numerical scores (and normalize them to between 0 and 1) and to model the uncertainties, we vary the value of each of the scores by a random amount. For example, a simulated person with poor physical health might have a Likert scale value of '2' (out of 5) to represent poor health. This is first normalized to 0.4 to fit within the range 0 to 1. A random value is drawn from a random normal distribution with mean 0 and standard deviation 0.1 is then added to this normalized value, for example, 0.4 - 0.0668 = 0.3332.

Once each attribute has been scored and a random factor applied, the ruleset is used to assign a decision to each applicant in the synthetic population. These decisions simulate the assessments that social workers make. The interpretation of the random variations applied to the attributes is that they represent individual variations in the judgements of the social workers. As seen in the game, the same applicants may be scored in slightly different ways by different social workers. Consequently, assessors may arrive at different assessments for the same individual. As a result of the addition of random variations, the same is true for assessments carried out using the ruleset on the synthetic population. A different set of random variations may result in a different assessment.

Our scoring exercise based on the social assessment ruleset of the existing system produces an output database with some individuals getting service and others not. We do not have the real-world socio-demographic data of social service recipients to validate the outcome in the Spanish case. Instead, we can check whether the database reproduces the number of people having received service in the real world (ca. 900,000 individuals), whether it reproduces the stylized facts on bias and discrimination in the literature [23], and whether it reproduces the case study's empirical research results.

5.2.1 Two examples for changing the ruleset towards the desired assessment system

In the game summary at the end of section 3, there are two ways in which stakeholders would like to see the current assessment system changed. To model the insight that an important attribute is missing from the SSM-Cat, Spanish citizenship, we add another attribute, 'citizen' for each member of the synthetic population and use this in addition for the scoring. To model the proposal that assessment should be the result of collaborative decision-making, the process of randomizing the agent attribute values and using the ruleset on these values to generate a decision is repeated various times (once for each simulated social worker) and the majority verdict is recorded as the final assessment for that individual. For example, if three social workers collaborate in assessing an individual, this would be simulated by using the ruleset three times for the same synthetic individual but using different random additions to the normalized scores each time. The effect of the random variations may be that the simulated assessments differ between each assessment, even though the ruleset remains the same. The final assessment would then be taken to be the majority decision (e.g. if two decisions allocate a budget and one not, the final decision would be to allocate).

Our new scoring exercise based on the social assessment ruleset of the desired system (generated by the iterative process between ABM and games) also produces an output database with some individuals getting service and others not.

5.3 AI-based social assessment

The assessment simulation on the synthetic population produces a dataset which contains one row for each individual, and where the columns represent the individual's self-sufficiency attributes as well as the amount of social services allocated to them as a result: full, partial or no social service allocation. The next task is for the machine learning (ML) component to learn the connection between the self-sufficiency attributes and the decision. Since each individual can have exactly one result and the 3 possible outcomes are mutually exclusive, a multiclass classification algorithm is required. The Keras library in the Python programming language was chosen to implement the AI component, as it is a widely used tool for ML tasks and has an active and large developer community. Its KerasClassifier class is suitable for the multiclass classification problem, and it uses a neural network-based approach. As these networks work with numeric input and output values, the desired output column needs to be transformed. One-hot encoding is applied on this column, which is a standard procedure to convert categorical values as input features for neural networks. As a result, instead of 1, the dataset now has 3 binary columns, one for each possible category.

After the dataset is prepared, the neural network is configured. For the size of the input and output layers, 6 and 3 were chosen to match the dataset. For the activation function in the output layer, "softmax" was selected as it renders a result where the output values add up to 1, which is required when there are mutually exclusive output categories. Choosing the number and size of hidden layers requires experimenting. In the end, one hidden layer with two times the size of the input layer seemed to be optimal for this dataset.

As a result of this, we have a ML model which can be provided with a set of personal attributes and which can generate an assessment decision that matches what a social worker would be likely to propose. For example, if the simulated dataset mimicking a real-world social worker were to show that severely disabled people receive a higher amount of social service provision, the machine learning component could learn this from the dataset and mimic the same or very similar decision for a self-sufficiency matrix of an unknown individual who is also severely disabled.

This process of simulating the assessment and training a ML component on the data can be performed on any ruleset suggested by the ABM interaction with serious games. Results can then be compared to each other in terms of fairness, for example by selecting a group of individuals that were experiencing bias in one system, and feeding their data into another trained on a different ruleset to see if it reduces that bias.

6 CONCLUSIONS AND OUTLOOK

Our participatory approach, combining serious games and ABM for stakeholder-driven innovation, led to several suggestions for better assessment practices provided by experienced social workers. Of course, there are certain limitations, for example, that only a small number of social workers can contribute their views towards system improvements. However, the general assessment approach in this specific national welfare system, the cultural concepts used, and the expertise perspective of professionals working in the field are sufficiently well represented.

In welfare systems, including in our Spanish case study, decision-makers often face complex cases that require a nuanced understanding of policies, individual circumstances, and broader social contexts. As a result, the process of making judgments about applicants' needs and eligibility is not solely an individual endeavor but also a collective one, where different players come together to discuss, deliberate, and reach consensus. This collaborative approach is reflected in our Spanish example and demonstrates a culture of interpretation where social workers or practitioners share insights, experiences, and perspectives to better understand and apply the rules and criteria governing welfare assistance. The use of methods that facilitate a collaborative environment in this way allows for the exchange of knowledge and experiences, leading to more informed and effective decision-making.

To test the reasonable suggestions of social workers familiar with the system, however, was not previously an option – especially not with a view to determining whether and how an AI system could take up these suggestions and make things better than they are. With the workflow described in previous sections, we have outlined a method for prototyping AI social assessment systems. Social workers can now see how their suggestions might affect the welfare landscape if implemented.

Next, we are planning a participatory evaluation where we will back to the participants of the game to ask them to check on the adequacy of the test assessments and on the effectiveness of the AI system were it to be put into a real environment. Incorporating an external, qualitative validation into the evaluation process of a welfare decision-making model presents a unique opportunity for stakeholders. This approach could significantly enhance the credibility, acceptability, and practical relevance of the model within the decision-making process in particular and the welfare system in general. We believe that stakeholders, including policymakers, social workers, and welfare administrators, might view this opportunity with optimism for several reasons. An external validation that incorporates qualitative assessments could provide a more holistic evaluation of the model by considering factors beyond mere numerical accuracy, such as fairness, ethical considerations, and the model's adaptability to complex individual circumstances. This could reassure stakeholders that the model not only performs well statistically but also aligns with the nuanced and human-centric nature of welfare provision. However, communicating this opportunity to stakeholders would require a careful and transparent approach. It would be important to emphasize the value of integrating qualitative insights with quantitative performance metrics to create a more robust and sensitive decision-making tool. Highlighting successful case studies or pilot projects where similar approaches have led to improved outcomes could also be persuasive.

Engaging stakeholders through workshops or seminars could provide a platform for direct dialogue, addressing concerns, and collaboratively exploring the potential of the model. From an institutional perspective, the external validation process could be located within a framework of ongoing quality assurance and improvement practices. It could be overseen by a dedicated task force or committee that includes a diverse range of stakeholders to ensure a broad perspective. This group could also be responsible for integrating feedback from the validation process into continuous model refinement.

Besides offering a useful tool to help social workers in reflecting and improving assessment behavior in their immediate workplace, the proposed approach is also relevant more generally for public, social and technology policy. Prototyping helps to avoid risk of failure, unintended consequences, and systems that turn out to be ineffective following expensive development. The option to address what-if questions, to test interventions before implementing them, and to evaluate the advantages and disadvantages of different scenarios, is of high relevance in many policy domains. Results will be discussed in the near future with representatives of the policy community responsible for AI use in public social service provision in Spain and elsewhere.

ACKNOWLEDGMENTS

This work was partially funded by the German VolkswagenStiftung under grant agreement number 98 560.

REFERENCES

[1] J. Angwin, J. Larson, S. Mattu, and L. Kirchner, "Machine bias," in *Ethics of Data and Analytics*. Auerbach Publications, 2022, pp. 254–264.

[2] V. Eubanks, *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. St. Martin's Publishing Group, 2018. [Online]. Available: https://books.google.de/books?id=pn4pDwAAQBAJ

[3] A. Alesina and G.-M. Angeletos, "Fairness and Redistribution," *The American Economic Review*, vol. 95, no. 4, pp. 960–980, 2005. [Online]. Available: http://www.jstor.org/stable/4132701

[4] L. Schwettmann, *Social Choice and Welfare*, vol. 38, no. 1, pp. 181–185, 2012. [Online]. Available: http://www.jstor.org/stable/41410220

[5] P. Taylor-Gooby and R. Martin, "Fairness, Equality and Legitimacy: A Qualitative Comparative Study of Germany and the UK," *Social Policy & Administration*, vol. 44, no. 1, pp. 85–103, 2010. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9515.2009.00701.x

[6] P. Ahrweiler, N. Gilbert, M. Bicket, A. Sabater Coll, B. Luque Capellas, D. Wurster, J. Siqueiros, and E. Späth, "Gamification and Simulation for Innovation," in *Advances in Social Simulation. Proceedings of the 18th Social Simulation Conference, Glasgow, UK, 4-8 September 2023*, C. Elsenbroich and H. Verhagen, Eds. Springer Nature. Berlin, Heidelberg, New York, 2024.

[7] P. Barbrook-Johnson and A. Penn, *Systems Mapping: How to build and use causal models of systems*. Springer Nature, 2022.

[8] O. Barreteau, M. Antona, P. D'Aquino, S. Aubert, S. Boissau, F. Bousquet, W. Daré, M. Etienne, C. Le Page, R. Mathevet *et al.*, "Our companion modelling approach," *Journal of Artificial Societies and Social Simulation*, 2003. [Online]. Available: https://jasss.soc.surrey.ac.uk/6/2/1.html

[9] T. Szczepanska, P. Antosz, J. O. Berndt, M. Borit, E. Chattoe-Brown, S. Mehryar, R. Meyer, S. Onggo, and H. Verhagen, "GAM on! Six ways to explore social complexity by combining games and agent-based models," *International Journal of Social Research Methodology*, vol. 25, no. 4, pp. 541–555, 2022. [Online]. Available: https://doi.org/10.1080/13645579.2022.2050119

[10] S. Strahringer and C. Leyh, *Gamification und Serious Games: Grundlagen, Vorgehen und Anwendungen*, ser. Edition HMD. Springer Fachmedien Wiesbaden, 2017. [Online]. Available: https://books.google.de/books?id=KeicDgAAQBAJ

[11] A. Caforio, A. Pollini, A. Filograna, A. Passani, and S. Filograna, "Design issues in Human-centered AI for Marginalized People," in *ITAIS 2021 Proceedings*, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:249579905

[12] A. M. B. Duarte, N. Brendel, A. Degbelo, and C. Kray, "Participatory Design and Participatory Research: An HCI Case Study with Young Forced Migrants," *Association for Computing Machinery Transactions on Computer-Human Interaction*, vol. 25, no. 1, feb 2018. [Online]. Available: https://doi.org/10.1145/3145472

[13] S. Lauriks, M. de Wit, M. Buster, T. Fassaert, R. van Wifferen, and N. Klazinga, "The use of the Dutch Self-Sufficiency Matrix (SSM-D) to inform allocation decisions to public mental health care for homeless people," *Community Mental Health Journal*, vol. 50, no. 7, pp. 870–878, 2014.

[14] M. K. Richmond, F. C. Pampel, F. Zarcula, V. Howey, and B. McChesney, "Reliability of the Colorado Family Support Assessment: A Self-Sufficiency Matrix for Families," *Research on Social Work Practice*, vol. 27, no. 6, pp. 695–703, 2017. [Online]. Available: https://doi.org/10.1177/1049731515596072

[15] Ö. Dilaver and N. Gilbert, "Unpacking a Black Box: A Conceptual Anatomy Framework for Agent-Based Social Simulation Models," *Journal of Artificial Societies and Social Simulation*, vol. 26, no. 1, Jan. 2023. [Online]. Available: https://www.jasss.org/26/1/4.html

[16] P. Ahrweiler, Ed., *Artificial Intelligence and Public Social Goods Eligibility Assessment - Achieving Social Justice Through Technology*. Springer Nature. Berlin, Heidelberg, New York, 2024, forthcoming.

[17] European Comission. (2021) Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52021PC0206

[18] T. Vihalemm, M. Männiste, A. Trumm, and M. Solvak, "Specialists and Algorithms: Implementation of AI in the delivery of unemployment services in Estonia," in *Artificial Intelligence and Public Social Goods Eligibility Assessment - Achieving Social Justice Through Technology*, P. Ahrweiler,

Ed. Springer Nature. Berlin, Heidelberg, New York, 2024, forthcoming.

[19] A. Sabater Coll, B. López, R. Campdepadrós, and C. Sánchez, "Participatory Action Research for AI in Social Services: An Example of Local Practice in Spain," in *Artificial Intelligence and Public Social Goods Eligibility Assessment - Achieving Social Justice Through Technology*, P. Ahrweiler, Ed. Springer Nature. Berlin, Heidelberg, New York, 2024, forthcoming.

[20] Institut d'Estadística de Catalunya. (2024) Idescat. [Online]. Available: https://www.idescat.cat

[21] Agència de l'Habitatge de Catalunya. (2016) Quantificació i distribució territorial de la població mal allotjada a Catalunya. Informe de resultats. [Online]. Available: https://img.arrelsfundacio.org/wp-content/pdf/AltresDocuments/2016_MalAllotjament_Catalunya.pdf

[22] Generalitat de Catalunya, Departament de Drets Socials. (2020) Informe sobre l'estat dels serveis socials a Catalunya 2020. [Online]. Available: https://dretssocials.gencat.cat/web/.content/03ambits_tematics/15serveissocials/sistema_catala_serveis_socials/documents/informe_estat_serveis_socials/Informe-sobre-lestat-dels-serveis-socials-2020.pdf

[23] Fundación FOESSA and Cáritas. (2022) Informe sobre exclusión y desarrollo social en Cataluña. Resultados de la encuesta sobre integración y necesidades sociales 2021. [Online]. Available: https://www.foessa.es/main-files/uploads/sites/16/2022/02/Informes-territoriales-2022-Catalu~na.pdf

AUTHOR BIOGRAPHIES

PROF. DR. PETRA AHRWEILER is Full Professor of Sociology of Technology and Innovation, Social Simulation, at Johannes Gutenberg University Mainz, Germany since 2013. Before 2013, she had been Full Professor of Technology and Innovation Management at Michael Smurfit School of Business, University College Dublin, Ireland, and Director of its Innovation Research Unit IRU. Furthermore, she was Research Fellow of the Engineering Systems Division at Massachusetts Institute of Technology (MIT), Cambridge/USA. Until 2022 she was the president of the European Social Simulation Association (ESSA). She studied Social Sciences at the University of Hamburg, Germany. At Free University Berlin, Germany, she received her PhD for a study on Artificial Intelligence, and got her habilitation at the University of Bielefeld, Germany, for a study on simulation in Science and Technology Studies. Her email address is petra.ahrweiler@uni-mainz.de.

PROF. NIGEL GILBERT is Professor of Sociology at the University of Surrey, UK, and is the Director of the Centre for the Evaluation of Complexity Across the Nexus (CECAN), which develops and tests methods for the evaluation of complex public policies. He founded and is Director of the Centre for Research in Social Simulation at the University of Surrey. The Centre has contributed new knowledge in a wide range of areas at the interface between engineering, public policy and the social sciences. He is also the founder and Director of CECAN Ltd, a spin-off from the research centre. He was one of the first to use agent-based models in the social sciences, in the early 1990s, and has since published widely on the methodology underlying computer modelling, and on the application of simulation for applied and policy related problems such as understanding commercial innovation, managing environmental resources such as energy and water, and supporting public policy decision-making. His email address is n.gilbert@surrey.ac.uk.

ZSOLT JURÁNYI, BSC is a lead software developer at DeepData Ltd. and at the Ethology Department of Eötvös Loránd University, Faculty of Science, since 2017. He holds a bachelor's degree in computer science. Between 2012 and 2017, he worked as a lead software developer at PetaByte Research Ltd., where his work consisted of web crawling, data analysis, web development and creating interactive visualisations. He was also involved in the national project FuturICT in 2013 and 2014, with similar responsibilities. In his work at DeepData, he mainly creates and maintains open-source full-stack web applications. The main focus of his work at the Ethology Department is mobile application development. His email address is juranyi@deepdata.hu.

MARTHA BICKET, MSC is a Senior Research Fellow at the University of Surrey and a member of the Centre for Evaluating Complexity across the Nexus (CECAN), specialising in the management and evaluation of complex systems. She has over 12 years‘ experience at the interface of research and policy, working closely with senior policy-makers and analysts across local, national and international levels of government towards the goal of better-informed decision-making and policy. She is one of the authors of the UK Government's central guidance on Handling Complexity in Policy Evaluation. Her email address is m.bicket@surrey.ac.uk.

PROF. DR. ALBERT SABATER COLL is Serra Húnter Professor of Sociology at the Faculty of Economic and Business Sciences of the University of Girona and Director of the Observatory for Ethics in Artificial Intelligence of Catalonia (OEIAC in Catalan). He holds a PhD in Social Statistics from the University of Manchester (UK) and since the publication of his doctoral thesis on the treatment of population data biases, he has worked on several research projects and collaborated extensively outside academia to tackle sociospatial and digital inequalities using a range of methodological approaches from social and computational sciences. He is a member of the Human Rights and Health Committee of the Generalitat of Catalonia, of the Advisory Council on Artificial Intelligence, Ethics and Digital Rights of the Barcelona City Council and of the Ethics Board of the European Lighthouse on Secure and Safe AI to advance research methodologies for safe artificial intelligence. He is currently leading the Spanish case study of the international project Artificial Intelligence for Assessment (AI FORA) and member of the leading team for the National Plan of Protection of Vulnerable Collectives in AI funded by the Spanish Ministry of Foreign Affairs and Digital Transformation. His email is email adress is albert.sabater@udg.edu.

PROF. DR. GEORGE KAMPIS is founding head (1994-2016) of the department of History and Philosophy of Science at Eötvös University in Budapest http://hps.elte.hu. After the department has been discontinued, he has been (since 2016) a Professor of Philosophy of Science in the Dept. of Ethology, to move (in 2023) to the Dept. of Artificial Intelligence. George holds a PhD and a Habilitation in Biology and a D.Sc. in Philosophy of Science. Main research interests in Artificial Life, cognitive science (Director of the Budapest Semester in Cognitive Science, http://hps.elte.hu/BSCS), evolutionary modeling and complex systems. He was a guest professor at Hokkaido University, Fujitsu Chair of Complex Systems at JAIST (Japan Advanced Institute for Science and Technology, in 2002/3), Wayne G. Basler Chair of Excellence at East Tennessee State University in 2007, and Fulbright Fellow at Indiana University (Bloomington) in 2009. He has been a project management associate at ETH Zürich on the (www.futurict.eu) EC flagship pilot. George works at the DFKI (the German Research Institute for Artificial Intelligence, http://www.dfki.de/web as a part time Senior Researcher in Embedded Intelligence since 2012. His email is kampis.george@gmail.com.

BLANCA LUQUE CAPELLAS, M has collaborated at the Sociology of Technology and Innovation, Social Simulation, at Johannes Gutenberg University Mainz, Germany since October 2021, where she is a research associate and is developing her PhD. She studied Sociology and Mediation, both in Barcelona. Between 2006 and 2020, she worked as a research assistant and project manager at several universities and institutions in Catalonia and Chile, where she also studied and trained in conflict mediation. Her research interests focus on the interrelation between society and technological innovations, social inequalities, research methods, and religion and society. Her email address is bluqueca@uni-mainz.de.

DAVID WURSTER, MSC is a research associate and PhD student at the Technology and Innovation, Social Simulation Lab at the Johannes Gutenberg University Mainz (Germany), since September 2022. He holds a bachelor's degree in industrial engineering (BEng) and a master's degree in Business Consulting and Digital Management (MSc). In his work he focuses on the ethical, social, economic, cultural, and environmental impacts of Artificial Intelligence, with particular interest on urban areas/societies. He is also part of the Artificial Intelligence for Assessment project (AI FORA). His email address is dwurster@uni-mainz.de.