Sandboxing Sustainable Tourism

Pablo Noriega[1], Manel Poch [2], and José Antonio Donaire[3]

[1] IIIA-CSIC (AI Research Institute of the Spanish Scientific Research Council), Barcelona, Spain
pablo@iiia.csic.es
[2] LEQUIA (Institute of the Environment), Universitat de Girona, Girona, Spain
manelpoch@fdasfgsdaf
[3] INSETUR (Tourism Research Institute), Universitat de Girona, Girona, Spain
jdoantiresadfas@iojoidsajfods

Abstract. Artificial Intelligence can contribute to sustainability in several ways. In this paper we focus on sustainable tourism and in the use of AI to support policy design. More specifically, we show how to design "values-driven" sandboxes for the design of sustainable tourism policies. These sandboxes are based on agent-based simulation and incorporate stakeholders' values in the modelling of their decision-making processes, as well as in the procedures of identifying policy ends, instrumenting the corresponding means and in the assessment of the results. We argue that the judicious use of sandboxes based on these ideas may support the deployment of actual policies with substantial effects on sustainability. We note that values-based sandboxes are also adequate research tools —in fields like tourism and sustainability— and can be framed as a paradigmatic problem for responsible AI research and development.

Keywords: values in AI, values-based sandboxes, agent-based simulation, sustainable tourism [4]

1 Introduction

Tourism has a profound impact on sustainability, accounting for roughly 3% of global GDP and one in ten jobs worldwide. With over 1.25 billion tourist trips recorded annually, the sector's influence extends beyond economic metrics, significantly affecting natural resources and social dynamics. For example, a tourist in a "sun and beach" destination consumes about 2.5 times more water per day than a local resident, and the arrival of a large cruise ship can severely disrupt local traffic even in well-developed seaside destinations.

Artificial Intelligence (AI) can improve sustainable tourism in several ways. Firstly, introducing AI-enabled tools and processes to bolster sustainable practices, such as placing smart sensors to monitor water usage and waste in residences, hotels and sports facilities, and managing traffic flow through intelligent control of traffic lights and access points. These AI applications can evolve into sophisticated, hybrid systems that

[4] UN SDG: G11: Sustainable cities and communities, G6: Clean Water and sanitation.

combine human and AI capabilities. For example, the concept of a smart city leverages data from these sensors to efficiently manage water supply and treatment, and adapting transportation and other services to the real-time needs of its residents. In turn, businesses can use insights from this data to present personalized offers to tourists and improve the overall visitor experience.

We propose a different approach to applying Artificial Intelligence for sustainability: to build flexible AI-enabled sandboxes to support stakeholders' decisions during the design, negotiation, and updating processes of sustainable policies. We postulate that, if appropriately used, AI should potentiate the sustainability effects of policies that are designed using an AI-enabled sandbox. The AI-enablement we advocate is the use of agent-based modelling as the kernel of the sandbox and the engineering of values into the sandbox in order to properly capture the multiplicity of interests and motivations involved in policy-making.

In this paper we use the particular case of a tourist destination to motivate and illustrate how to design such a values-driven sandbox and how it can be used to identify a sustainability policies that are (i) effective in achieving consensual sustainability targets, (ii) implemented through adequate means-ends trade-offs and (iii) acceptable to all stakeholders. The case of a tourism destination is general enough to describe the essence of our approach, yet specific enough to demonstrate how to implement a values-driven agent-based policy-design sandboxes for different scales and for other policy domains based on this approach.

The paper is organised as follows: Section 2 outlines the main tenants of our proposal and gives a bird's eye view of its motivation and background. Sec. 3 is a succinct description of the functionality of a sandbox and how the core components of the policy domain are modelled. Sec. 4 explains how the different types of stakeholders are brought into the sandbox design and Sec.5, how to model values and to assess how successful a set of policy means is in achieving the postulated policy ends. We sketch the way the sandbox is used to identify deployable policies in Sec. 6 and we sum our argument up in nine concluding remarks .

2 Background

2.1 Modelling policies for sustainable tourism

In raw terms, a policy is a way of provoking changes —usually through infrastructure development, regulations, incentives and community engagement— in the state of the world in order to reach a better one. In this paper the world to be improved —the *policy domain*— is a *tourist destination* and the frame of reference is *sustainability*.

Tourism impacts the sustainability of a destination in three key dimensions: environmental, economic, and social. Environmentally, tourism affects water consumption, energy use, waste generation, and CO_2 emissions, which can harm local ecosystems, and high-density tourism, for example, can lead to overcrowding and strain on infrastructure and resources. Economically, tourism influences local economies, potentially bringing both benefits like added income as well as challenges such as dependency or inequality. Socially, tourism impacts are reflected in the perceptions and satisfaction of both tourists and residents, affecting community well-being and cultural preservation.

In this context, municipal and regional policies are meant to play a crucial role in achieving a sustainable development of a tourist destinations. However, postulating a potential policy and foreseeing its effects is a challenging problem due, intrinsically, to the uncertain and complex interplay of multiple stakeholders and how these affect and are affected by the local environment of the tourist destination, as well as the wider socio-cultural-economic environment where that specific destination is situated.

That type of complexity is not specific to sustainable tourism and is the core of of the *policy-design process* in general (see for example, [3,5,29]). Moreover, given the significant social and economic impact that the deployment of a policy can have, the need of an *ex-ante impact assessment* is a recommended practice at the heart of policy design and simulation is a prominent resource towards that end [7,4]. In fact, it is acknowledged that because of the large number of variables and the dynamic nature of their interactions, policy design —in a tourist destination as well as other policy domains— is what is known as a "wicked problem" involving factual and ethical decisions (cf. H. Simon [28]). And, as is the case with other problems of this sort, simulation is arguably a reasonable methodological approach to policy design [1,9,12,22].

In the case of policy design and impact assessment, the primary requirement of simulation is to produce realistic predictions to assess whether such policy is well-suited for its intended purpose (see [11]). Nevertheless, simulation can serve other purposes [6] and misuses have been aptly noticed ([2]). Our understanding is that the use of a simulation in a policy design sandbox should serve not only *epistemic* (produce realistic predictions) but *rhetorical* purposes (it should be informative in order to gain support) as well.

The usefulness of a simulation depends on the availability of a good enough model and the repertoire of modelling techniques is large and well understood. However, for the purpose of policy design, traditional models often overlook the role of individual decision-making processes and accounting for the different needs and preferences of stakeholders and the potential inherent conflict.

Agent-based simulation addresses the first issue ([10]). Agent-based simulation is based on the premise that some complex social phenomena emerge as the result of the collective behaviour of individuals. Its distinctive advantage is to model both the decision-making of individuals as well as the coordination mechanisms that constrain and enable their collective activity.

The second issue is addressed by bringing values into the simulation. While agent based simulation adds the possibility of modelling social aspects on top of more conventional engineering or econometric techniques, the explicit modelling of values into the decision-making models of individuals, the policy domain and the definition and assessment of policy interventions, allow the analysis and proper representation of the expectations, motivations and interests of the different stakeholders.

It goes without saying that, in order to build robust and reliable simulation models, policy domain models as well as individual decision-making models need to be supported with conventional forms of modelling, solid data sources and awareness of the political environment.

2.2 Sandboxes for policy-making

Broadly speaking, sandboxes have two main advantages as a policy design tool: (i) *Epistemic*: as a means to understand the policy domain and visualise the impacts that policy interventions would have, and (ii) *rhetorical*: as a communication device to facilitate the stakeholders to agree on a policy proposal.

Sandboxes, as interactive design workbenches, help understand the complex nature of the policy domain and potential interventions. A salient advantage is bringing into light unforeseen consequences and assess their significance and cost. In this respect, they are particularly useful for exploring unconventional means to achieve policy goals and find out what their outcomes would be.

Sandboxes implement a robust form of evidence-based assessment of policy proposals, thus from a *rhetorical* perspective facilitate consensus among stakeholders. In particular, having stakeholders values engineered into the tourist destination model is a powerful way of elucidating trade-offs and equilibria among design stakeholders, facilitates consensus building, hence the negotiation and approval of a consensual policy.

As a convenient side-effect, once the design phase is over, sandboxes may be used to support deployment and follow-up of the agreed upon policy. Namely, any discrepancy between forecasted parameters and their score while the policy is being enacted rises the opportunity for agile reaction. A discrepancy could either point to inadequate forecasting —a weakness of the model or an unforeseen deviation from the assumptions about the pertinent starting conditions or about their evolution— hence, to the need of updating the modelling assumptions, or point to a misapplication of the agreed upon policy instruments —hence, the occasion for a revision of the instruments, or a correction of the unwanted effects of their misapplication.

2.3 Values-driven sandboxes

While agent based simulation adds the possibility of modelling social aspects on top of more conventional engineering or econometric techniques. However, modelling values for policy design is not straightforward. Values serve four different functions in the sandboxes we propose: (i) values determine what is the state of the world that is meant to be improved and what an improvement means; (ii) values determine how to compare states of the world and decide which one is preferable; (iii) values determine which means are preferable or acceptable to change the state of the world, and (iv) values guide the behaviour of simulated agents.

While the modelling of values captures differences and conflicts of interests among stakeholders, they are ”engineered” into the model also to reflect some degree of consensus among the stakeholders who are involved in the modelling and the validation of results of the use of the sandbox. Because if the sandbox is used to negotiate a policy, the modelling of values establishes the “courses of action” that are admissible: values identify and ultimately legitimise the chosen policy goals, values help identify the means (affordances and constraints) that can be used to modulate the behaviour of participating agents towards achieving the chosen goals and values provide an objective way to tell which potential policy interventions may be worth deploying (Sec: 5.

In line with these considerations we also propose not only to engineer stakeholders' concerns into the software agents that simulate stakeholders at the core of the simulator, but also to bring involve the actual stakeholders into the process of designing —and using– the sandbox. In particular. we are proposing a *participatory design process* that is based on the need to reach consensus among these stakeholders, since the purpose of a sandbox if to choose a policy that is to be deployed with their support (Sec. 6). Our proposal supports such consensus with two main devices:

The first device is how values are engineered into the sandbox (Sec. 5): In order to avoid a discussion about values themselves, we look for "goals" that stand for those values. Thus, value engineering is geared towards reaching consensus around concrete indicators that are found to be compatible or legitimised by stakeholders' values, and on the preferred means to reach them. Once an acceptable set of indicators and the means or instruments to reach them are tentatively agreed upon, stakeholders need to reach an agreement on the way to tell how satisfactory the application of those instruments is. That part is made operational through the specification of the *assessment functions*. An *effectiveness* assessment function captures the agreement of stakeholders on the relative importance of these indicators and the degree to which they need to be achieved. The specification of the *adequacy* assessment function establishes a consensual cost-benefit analysis between alternative sets of instruments to reach the consensual goals; and the *acceptance* assessment function captures the cost-benefit analysis of those same instruments but with respect to each individual stakeholder's own cost-benefit assessment.

The second device is to let stakeholders' views be expressed at different "levels of abstraction" (Sec. 4). First into the validation of the simulated stakeholders decision-making models that reflect what aspects of the domain each one is concerned with, what are the main motivations and how these are best served with different states of the world. Second by their involvement in the definition of indicators, targets, cost-benefit analysis, specification of the three assessment functions; and in their involvement in the calibration and validation process of the simulation model (Sec. 5). Third, in their involvement in the use of the sandbox for the definition of likely policy interventions and finally in the choice of a policy (Sec. 6).

2.4 A wider view of values-driven sandboxes

The sandbox we present in this paper is intended as an illustration of values-driven sandboxes. Although we use tourism to illustrate our points, the modelling of an actual tourism destination requires a detailed specification of all the components sketched and illustrated below. Nevertheless, the structure and methodological heuristics we present here can be used as a blueprint for sandboxes of other tourist destinations, other tourism policy scales and to other sustainability policies

Moreover, While the resulting (fully instantiated) sandbox would serve as a policy-design tool, it may also be used as a tool for an experimental approach to the study of tourism (for example, to account for paradoxical outcomes) and develop richer data spaces to reinforce the underlying models.

In addition, the construction of a policy design sandbox can also be taken up as a paradigmatic case of the so called value alignment problem in AI (VAP). That is, the design of artificial intelligence systems (AIS) that are objectively aligned with human

values [24,8,26,13] . A values-driven sandbox like the one we discuss in this paper can be used as a test-bed for different ways of formulating fundamental questions about the role of values in the control of artificial autonomy both at the level of an individual autonomous software agent or in the regulation of the collective activity of autonomous agents that might as well be natural or artificial [21,14,17,20]

3 The Design of a Sandbox for Sustainable Tourism

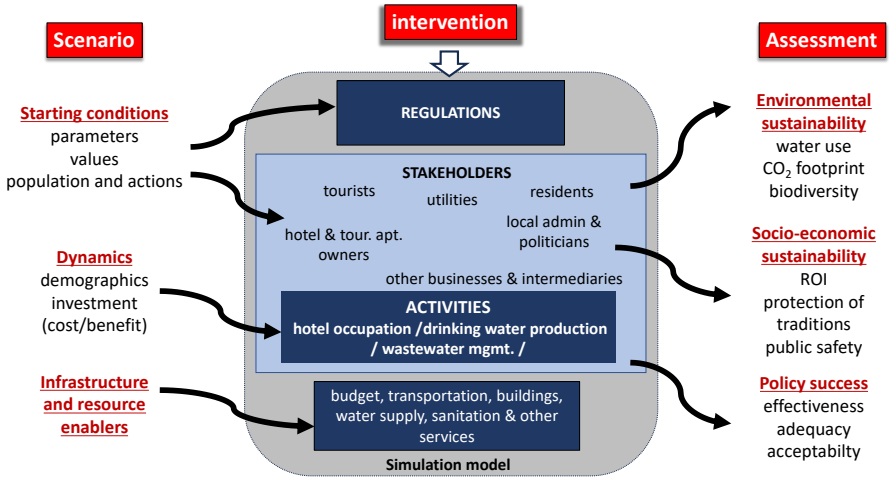

Fig. 1: A sandbox for sustainable tourism, is a simulation workbench that is used to assess the effects of policy interventions on sustainability variables of a tourism destination under different assumptions.

In this paper we outline how to build a sandbox for the design of sustainability policies in tourism. The sandbox is an interactive simulation environment based on an agent-based model of a *policy domain* that, for illustration purposes, in this case is a tourist destination.

Fig. 1 illustrates the relevant aspects of that policy domain and how the simulation process works.

The sandbox is put to work by loading the simulation model with a given *scenario* specified through some assumptions of what the starting conditions of the *state of the world* are at the moment when a simulation starts, and how these starting conditions are expected to evolve.

The activity of the population of simulated agents within the policy domain become activated with a *policy intervention* —that is, a set of policy means that govern the activity of the simulated stakeholders plus the criteria to evaluate how successful that intervention becomes (5). The effects of such activity on the state of the world becomes observable trough selected sustainability indicators.

As illustrated in Fig. 2, the simulation model consists of two components that capture the persistent aspects of the policy domain: a *physical model* (Φ), and an *institutional model* (Ψ). These two models —the physical model (Φ) and the institutional (Ψ)— are meant to be fixed when interventions are introduced into the simulator. They establish, respectively. the *natural* affordances and constraints of the model: that is, the way the model mirrors the real world and the *artificial* constraints and constraints that are added on top of the natural constraints in order to "articulate" simulated agent interactions ([18]). Interventions, which are the focus of the experimental use of the sandbox (Sec. 6), become part of the simulation model on top of these two "persistent" models by adding additional affordances and constraints.

The simulation model also includes a collection of simulated stakeholder models ($M = \bigcup_{\kappa} \mu_{\kappa}$) that account for the behaviour of individual stakeholders —like residents, hotel managers and owners, and tourists— that become active during simulation runs.

This model is specified as an online hybrid multiagent system (in fact, an "online institution" [17]) implemented on an agent-based simulation platform.

Since we are building a values-driven sandbox, special attention is given to the embedding of values in the modelling of the simulated agents (Sec. 4), and as we discuss in Sec. 5, values are also central for the definition of a policy intervention and the evaluation of its effects.

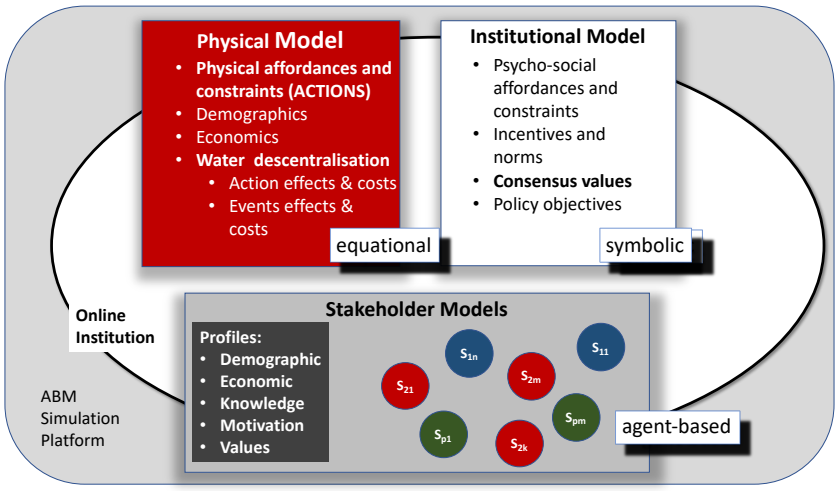

Fig. 2: Agent-based simulation of a tourism destination involves three complementary components that are implemented as a multiagent system on an agent-based simulation platform.

Ultimately, the sandbox is used to explore in a systematic way what are the preferred policy interventions in order to decide whether the set of means included in a given

policy intervention are worth actually implementing or not. This exploration process outlined is sketched in Fig. 7.

3.1 Modelling the persistent aspects of the policy domain.

The physical model (Φ) is an abstract representation of *the relevant part of the world*. That is, a model of the policy domain that includes those entities of the "physical" reality that support or affect the activity of individuals in the policy domain and the features of that relevant part of the world that constrain or enable the evolution of the actual policy domain. Hence, Φ establishes: (i) those facts that constitute the *state of the world at time* t_i (in addition to all sustainability variables/parameters, the state of the world also contains additional facts or variables —like, the number of tourists in a given hotel at time t_i or the cost of purifying a ton of waste water— that are necessary to calculate these); (ii) those external events (a forest fire, exchange rate fluctuations) that can change the state of the world, the conditions to acknowledge their occurrence and their effects; (iii) the physical affordances and constraints on the actions that simulated stakeholders can take during a simulation (occupy a room in a three-star hotel and use its facilities at time t_i, invest in a water-treatment plant) together with the (physical) preconditions for and the consequences of their execution (notice that, as discussed below, persistent legal and procedural preconditions are contained in the institutional model).

In practice, sustainability variables are modelled mostly using conventional equational models —econometric, game-theoretic, mathematical programming— on assumptions about input parameters, initial conditions, and the effects of the successful actions that are compatible with a given set of policy instruments (and the artificial constraints established in Ψ. For example, we can model the actual water usage in three-star hotels by assuming a specific flow of tourists over the year, water usage per day and room, other hotel water uses (like house-cleaning, gardening, sport facilities, cooking), occupancy rate ad tourist profiles (budget, values, personality). The results can be visualized in graphs showing standard use, optimized use, and simulated use under different policy scenarios.

From a simulation perspective, Φ ought to be realistic, detailed and reliable enough to provide sound evidence about the outcomes of a proposed policy. Hence, it needs to make accurate predictions to measure effects of actions and events. It also needs to be informative enough so that policy-makers may understand the relevant modelling assumptions and evaluate the effects of policy interventions.

In practice, sustainability variables are modelled mostly using conventional modelling techniques on assumptions about input parameters, initial conditions, and the effects of the successful actions that are compatible with a given set of policy instruments. For example, we can model the actual water usage in three-star hotels by assuming a specific flow of tourists over the year, occupancy rates, budget constraints, value profiles of tourists, and the impact of incentives and motivational campaigns. The results can be visualized in graphs showing standard use, optimized use, and simulated use under different policy scenarios.

The institutional model (Ψ) While Φ models the physical constraints that affect agent interactions as well as those physical affordances that enable them to act, the institutional model, Ψ, captures "artificial" constraints and affordances. Constraints and affordances that are imposed on the activity of simulated agents or groups of agents.

For example, legal and administrative procedures that enable the availability and use of subsidies for retrofitting hotels with water-saving devices, as well as the regulations that require retrofitting; promotional campaigns to attract higher paying, longer staying tourists; awarding sustainability and proximity market recognitions.

Note that, in particular, the institutional model ought to capture the value-related aspects that concern the choice, interpretation, instrumentation and assessment of values in the design and use of the sandbox, as discussed in the next section.

4 Stakeholders in a tourist destination

Stakeholders are brought into the design of sandbox in three roles: policy *owners* (the ones who are responsible for enacting the policy), policy *designers* (the ones who are involved in the a design of the sandbox and in its use for exploring alternative policy interventions that would be fine-tuned by policy owners) and *simulated* stakeholders (the ones whose agent models are part of the simulation model outlined in Fig. 2).

4.1 Policy owners.

These are the actual individuals and institutions who use the sandbox to fine-tune policy proposals and eventually decide to enact one. They include (i) tourism authorities (local, regional and supra-regional); (ii) the politicians whose activity impinges or is directly affected by tourism; (iii) associations that speak for collective stakeholders like tourist councils, better business bureaux or associations of residents; and (iv) significant service providers like utility companies, large investors or financial services.

In addition to their ultimate role in the definition and adoption of a policy, *policy owners* contribute input for modelling the policy domain and decide the principles for the evaluation of policy proposals.

4.2 Design stakeholders

These are those parties —individuals or teams of individuals– involved in the actual design process of the sandbox: these include, (i) the designers and experts who model the policy domain and build the sandbox, (ii) the policy maker who is in charge of articulating a policy proposal —in fact, the policy-maker determines the core components of the models, runs simulations with the sandbox to explore actual proposals—; and (iii) representatives of the policy owners who will participate in the construction of the model and the fine-tuning of the simulator.

Design stakeholders would participate off-line in the choice and interpretation of the values that bear upon the policy and they have to reach consensus not only on such choices and interpretations of values, but also on how these are reflected in the

selection of policy instruments and in the assessment of the success of the policy; i.e. design stakeholders run the simulation process and validate the outcomes.

4.3 Simulated stakeholders

These stakeholders are software agents that model persons, groups of persons and firms whose individual or collective activity affect the state of the world.

These stakeholders are simulated as autonomous software agents whose actions are subject to the constraints and affordances specified in the physical and institutional models. The activity of these simulated stakeholders is reflected in the evolution of the sustainability variables over a period of time. Since those actions are further modulated by the policy instruments that become active in a given policy intervention, the activity of those simulated agents provides the basis to evaluate and compare policy interventions when using the sandbox.

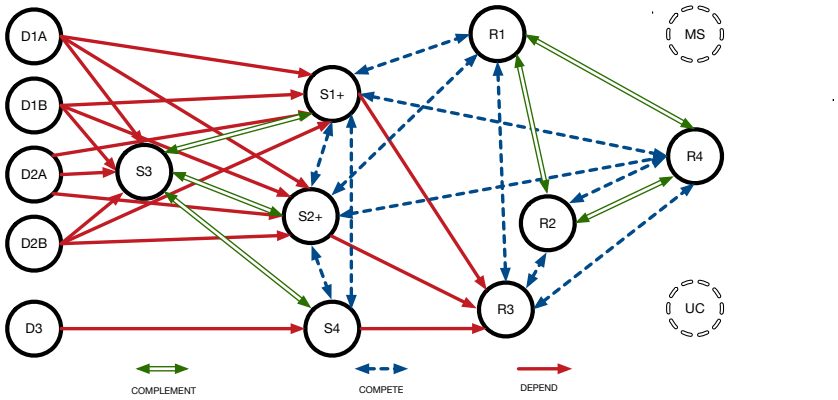

Fig. 3: Example of the main simulated stakeholders in a tourist destination and their interrelationships: types of tourists (D); hotels and intermediaries (S), and residents (R). Municipal services (MS) and utility companies (UC) stand for several stakeholder classes that might not be simulated.

For sustainable tourism sandboxing one may start with fifteen simulated stakeholder types outlined in Fig. 3 that can be organised in the following five classes:

1. *Demand stakeholders.* (D1) Tourists who respond to S3 offers, with a sustainability orientation: (D1A) or without (D1B); (D2) Loyalty tourists (who return to the destination), with a sustainability orientation: (D2A) or without (D2B); and (D3) tourists who prefer residential lodging over hotels.
2. *Tourism suppliers properly.* (S1+) local hotel owners: that is,family operated establishments whose owners are local residents (several categories) (S2+) corporate hotel owners: establishments that respond to corporative criteria and have little or no commitment to the destination (several categories). (S3) tourism intermediaries

(firms, like tour operators and travel agencies that connect supply and demand and may divert demand according to local conditions); and (S4) residential lodging intermediaries (like *Airbnb*).

3. *Resident stakeholders.* (R1) local residents whose income is not directly dependent on tourism; (R2) local residents whose economic activity takes place in environmentally or heritage-significant fields (agriculture, utility companies, transportation, religious and cultural orgs.); (R3) local residents whose income is directly dependent on tourism (waiters, hotel staff, taxi drivers, tourist guides,...); and (R4) local residents involved in environmental and heritage preservation.
4. *Municipal services.* (MS) Depending on the granularity of the model, entities that are part of the local, regional or supra-regional administrations and are involved in sustainability-related activities like security, social support, public transportation, parks, gardens and waste management. This class might also include simulated *policy-support and enforcement agents* that can gather and report emergent information to different stakeholders or stand for the type of decision-making that the tourist destination as a "smart city" could implement. These entities and activity may be simulated as autonomous agents of different types that can sense and act in the policy domain but, occasionally, they might be modelled into the agent model as functions or parameters.
5. *Utility companies.* (UC) Likewise, one might also include in the model other stakeholders that are not local residents or not as ostensibly linked to tourism as those in S3. Namely, utility companies and other non-public entities that are involved in sustainability related activities or provide sustainability related services not necessarily directly related with tourism, like water, communication and energy utility companies, waste processing plants, industry and neighbouring farms. Like municipals services, these stakeholders may be modelled as simulated agents, or simply as functions or parameters.

Simulated stakeholders, are modelled as autonomous software agents who can choose to take any "afforded action" —actions that are available to the agent and have an effect in the state of the world— when the relevant conditions for its execution are satisfied.

For policy simulation in particular, the salient stakeholder decisions to be modelled are those that activate an action that has an effect on the policy objectives (sustainability indicators). For example, a tourist that is environmentally conscious would decide to follow the suggestion to reuse hotel towels and loyalty tourists might prefer to stay in a tourist apartment and buy at the local grocer. A large tour operator would direct tourist flows only towards "group-friendly" destinations. Hotel owners would co-finance a new water treatment plant if, for example, water supply is expected to be insufficient in the foreseeable future and no cheaper sources are available; in that scenario, the local government might facilitate this decision through economic incentives and streamlined regulation.

In practice, simulated agents are implemented on top of agent models (μ_κ) that capture the repertoire of actions that are available to a class κ of agents whose behaviour should be properly differentiated. For instance, depending on the granularity of the model, simulated hotels of type (S1+) may include one or several classes (because 3-star hotels may have radically different capabilities and needs than 5-star ones). These agent

models (μ_κ) include two elements: the specification of the decision-making process that supports those actions that are affordable to the members of the class and the way that members of that class assess the degree to which a policy intervention satisfies its own objectives (σ_κ).

Likewise, the μ_κ decision processes are implemented through three main strategies (that correspond to three distinct forms of value-driven decision-making identified in behavioural economics and neuroeconomics, see e.g. [23]): reactive (instinctive), adaptive (learning, virtue) and rational (value-driven reasoning). As suggested above, depending on the granularity of the model, individual or collective stakeholders stakeholders in classes MS and UC may be implemented as simulated agents or implemented directly as affordances and constraints in Φ and Ψ. Moreover, some of these stakeholders may not be simulated but simply use of the sandbox as *design stakeholders*.

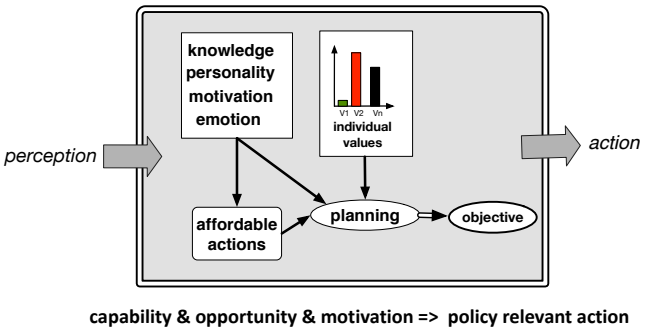

Fig. 4: An agent decision making model (μ_e). The actions that an autonomous agent takes depend on the state of the world it can perceive and a decision to choose an available action in accordance to its capabilities, goals, values and other cognitive constructs. (From [14])

For those policy-relevant actions, as Fig. 4 shows, agent modelling assumes that an agent takes one policy relevant action if and only if the agent is capable (has the entitlements, resources and ability), motivated (values and other cognitive features are involved in such motivation) and has the opportunity to perform it (is aware of the possibility and the conditions for execution are fulfilled).

The simulation —as Fig. 3 shows— should take into account not only the capabilities —and decision-making processes— of each of the fifteen classes of simulated agents but also the type of compatible and conflicting motivations. These features are reflected in the interpretation of values for each class of simulated stakeholders and captured in the respective decision-making processes. However, these adversarial features need to be addressed also in the institutional model and policy intervention instruments, in terms of affordances (residents may oppose the construction of new hotels), constraints (tour operators can block hotel rooms on a yearly basis but commit to actual occupancy four weeks in advance) and information (campaigns to motivate residents

to support environmental sustainability measures or smart sensors that warn hotel managers when their water use is above expectations, for instance).

5 Modelling and assessing policy interventions

As suggested above, the point of the sandbox is to design a policy by exploring in a systematic way what the objectives of the policy could be and identifying appropriate means to achieve those objectives (see Fig. 7). This exploration process is grounded on the possibility of simulating what are the actual effects of a particular combination of means and assessing how successful this combination is. In this section we discuss how we represent a policy intervention —the input that sets the simulation in motion— and how to "engineer values" into the sandbox in order to be able to assess how "successful" those means are.

5.1 Policy interventions

A *policy intervention* $\Pi =< I_n; \Sigma >$ is a specific set of policy instruments (I_n) and the elements to assess (Σ) how successful they are in achieving some policy goals. the systematic exploration process works.

As suggested in Fig.1, we start with a good enough model of the policy domain where a population of simulated stakeholders interact in accordance to the affordances and constraints provided by the physical and institutional models. The set of policy instruments I_n (means) whose effects we want to assess provide affordances and constraints on top of the ones already in the policy domain (in Φ and Ψ). Those policy intervention instruments can be of four sorts: (i) *new capabilities*, for instance, in the form of subsidies for water saving retrofitting; (ii) *new regulations* like the procedure and requirements for licensing a tourist apartment; (iii) *information* to motivate simulated stakeholders' behaviour like nudges to motivate water saving practices and promotional campaigns to attract longer pernoctations and higher paying guests; as well as (iv) *simulated policy support and enforcement agents*, like, for instance, smart sensors of water use, and aggregators and managers of smart city data (like stakeholders of type MS above).

We propose a three-fold way $\Sigma =< \sigma_e, \sigma_q, \sigma_{acc} >$ of assessing how successful those means are: ***e**ffectiveness* (with afunction σ_e, to assess to what degree policy goals are achieved); *ade**q**uacy* (with a function σ_q that aggregates how "suitable" is Π in terms of cost/benefit trade-offs between intended, unintended and unwanted consequences of the instruments in Π); and ***acc**eptability* (with a function σ_{acc} that, in fact, aggregates the degree to which values are satisfied for each of the simulated stakeholders (one aggregation function, ${\sigma^{\kappa}}_{acc}$, for each of the μ_κ classes of stakeholders simulated in M). Each of these three criteria is implemented with a "value-aggregation" function that depends on the set of values of the stakeholders and also on the way these values are represented in the model as a set of goals $\Gamma = \{< \gamma_i; t_i >: i = 1, ..., m\}$ (index+target pairs) explained below.

5.2 Values in policy making simulation.

A distinctive feature of our proposal is to incorporate values as a fundamental component of a policy design sandbox. Values motivate objectives to be the achieved and identify equilibria among stakeholders' interests in order to reach consensus and support of a given policy. Ultimately, as mentioned above, values are the key to assess how successful a policy intervention is and, therefore, to decide whether a particular policy intervention should be deployed or not.

This generic understanding of the role of values needs to be made operational in the sandbox. We propose to to "engineer" values into a sandbox for sustainable tourism in a a four step process that we describe next. A full account of this matter is beyond the scope of this paper, see f.i., [20,19,17,16] for a wider perspective.

Step 1. Value identification . We assume that *values motivate goals*, goals can be made explicit as *objectives*, which are made observable in the simulation as *indicators*. For example, Fig. 5 captures a fraction of the representation of the value "environmental sustainability" where one can postulate vague intuitive goals and move through a goal decomposition process to a set of parameters that help identify actions that affect the value in question.

In *stricto senso*, one should distinguish between an "indicator" —which stands for a parameter that is directly observable in the state of the world (in Φ), for example daily water use per room in three star hotels— and an "index" which combines several indicators into a single parameter —like the total reduction of water in the location. The choice has to do with the granularity of the model and the way several indicators or several indexes are aggregated in order to assess the success of a policy.

Step 2. Value interpretation. In order to identify policy means and assess their success, we assume that for every indicator that stands for a value, and at any step of the simulation, one can compute a score, which is interpreted as a degree of satisfaction of the corresponding value in that state of the world. In fact, we establish a *preference relationship* between the possible scores of each indicator —and a *target score* or "aspiration level"— that will be used to assess the degree of satisfaction of that value. We define the set Γ (for policy "goals") as the list of pairs that include, for each value V_i, its indicator γ_i (whose scores s_i are ordered) and its target t_i.

Step 3. Value instrumentation. Once values are associated as indicators, one can identify those actions that have an effect on those indicators. Since agent actions always have effects on the state of the world, those control mechanisms that either promote actions that contribute to the achievement or protection of a value (a positive effect on the corresponding indicator), or discourage or prevent those actions with unwanted consequences (decrease the score of the indicator) become the *means* to achieve the policy goals. In this way, institutional components and policy instruments like the ones we mentioned in Sec. 5, modulate simulated stakeholders' activity towards the goals of the policy.

Fig. 5 shows the *consensual interpretation* of the value *environmental sustainability* in its subcase "*sustainable water use*". Here, the interpretation of a value corresponds to an ordered list of policy goals all stakeholders agree upon, and each policy goal is then associated with some means to achieve it.

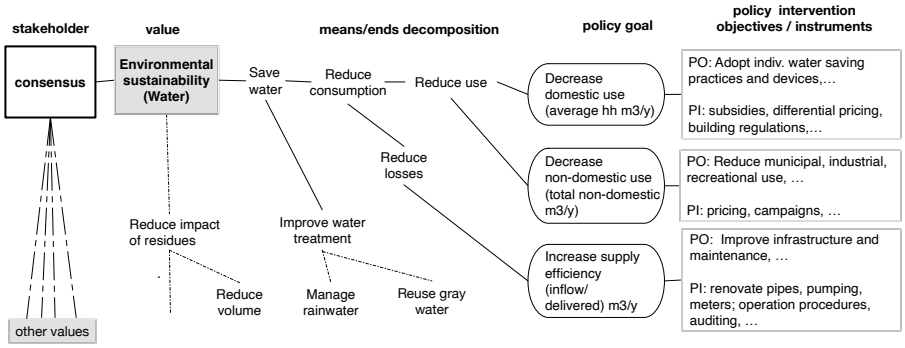

Fig. 5: From values to policy goals and instruments. The diagram shows water use indicators and potential instruments for the top part of a graph of the consensual interpretation of the value "environmental sustainability" (after [14]).

Note that, in the process of making values operational, one may need to update the definition of the physical model (Φ) in order to include those elements needed to represent the values (parameters, variables, indicators), the relevant actions that affect those indicators and the execution of actions involved in policy instruments. Likewise one needs to update the institutional model (Ψ) to accommodate the affordances and constraints that support the corresponding policy instruments.

Note also that since different stakeholders have different values, two stakeholders might end up with different interpretations. However, since any given policy intervention Π will apply to all stakeholders, all stakeholders ought to reach a consensus on the objectives worth reaching and the corresponding indicator-target pairs ($\Gamma = \{< \gamma_i, t_i >: i = 1, ..., m\}$). Not withstanding this agreement on goals, the way different stakeholders assess their satisfaction with the outcomes of a given intervention might differ. These different interpretations of the "successfulness" of a given intervention are addressed through the "acceptabiliy" criteria described below.

Step 4. Value assessment. In order to tell how "good" or how "successful" a policy intervention is, we propose to take three different perspectives: how successful it is in fulfilling the agreed upon goals, how it compares with other possible ways of fulfilling the same goals and how "satisfied" is each stakeholder with that intervention (and what is the collective satisfaction that results). For each of these perspectives we define a function that aggregates (or synthesises) the contribution of all the values to the corresponding understanding of "success" ($\Sigma = \{\sigma_e, \sigma_q, \sigma_{acc}\}$)

1. Effectiveness. An obvious way to measure the success of a policy intervention is in terms of its "*e*ffectiveness". Thus, σ_e; measures to what extent each and all of the policy goals (Γ) are met. In other words, σ_e is an aggregation of the degree of satisfaction of all the policy goals in Γ.

This aggregation –as well as the aggregations in σ_q and σ_{acc}— is some sort of multi-objective function that reflects the relative importance of each goal and how distant are the intended goal target and the actual score in the simulation.

This aggregation function can take several forms, for example, one can use an aggregation function —like the one in Fig. 6 inspired by H. Simon's "satisfycing" notion [27]— that calculates effectiveness as a weighted sum that reflects the relative importance of each value and the proximity of the actual score of a value with respect to the target "aspiration level" of that value.

Or more formally: If G calculates the importance of the gap between the score of value V_i in the simulation (s_i) and its target score (t_i), the effectiveness of an intervention Π

$$\sigma_{\mathbf{e}}(\Pi) = \sum_{i=1}^{n} G(s_i, t_i) \times \omega_i \text{ where } \omega_i = \frac{W(V_i)}{\sum_{i=1}^{n} W(V_i)}$$

is the weighted sum of the n values $V_1, ...V_n$, ; and ω_i is the normalised contribution of V_i ($W(V_i)$) to the aggregated satisfaction.

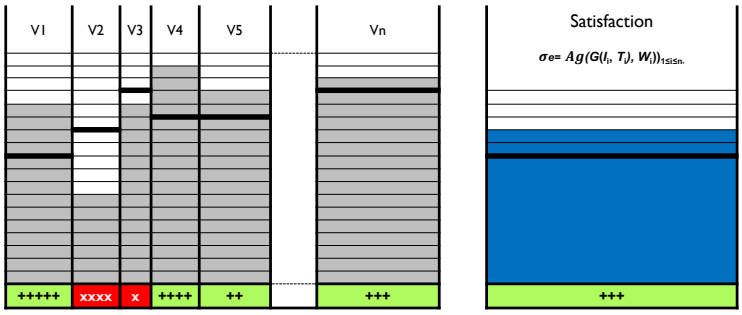

Fig. 6: A simple value aggregation *satisfycing* function. Each value in $(V_1, ..., V_n)$ has a target threshold, or *aspiration level* (dark line), a relative importance with respect to other values (width of the column) and relative fulfilment in a given state of the world (coloured cells). Their aggregated satisfaction is represented in the last column. In this case, the state of the world is "effectively aligned" with the set of values since the satisfaction with most values compensate the dissatisfaction with values V_2 and V_3 of lesser relative importance. (From [14]),

2. Adequacy. A second aggregation function (σ_q) measures success as the "ade*q*uacy" of the policy instruments in Π for achieving Γ. The point is to assess how appropriate or suited are the means that are used to achieve the goals. Intuitively, σ_q calculates what is the cost of the effectiveness of Π. Hence, the purpose of σ_q is to have a way of

comparing two policy interventions Π_1 and Π_2 —that are sort of equally effective with respect to the same goals— by determining the respective cost/benefit trade-offs. These trade-offs can be calculated by combining the costs and benefits incurred in achieving the intended objectives and the costs an benefits of producing unwanted as well as unintended outcomes.

3. Acceptance. Note that, both, σ_e and σ_q are calculated with respect to the consensus values but one should also assess how compatible a given policy intervention is with the respect to the values of each of the different stakeholders. This concern is addressed with the third function σ_{acc}

The acceptance function σ_{acc} is a combination of the degrees of effectiveness and adequacy that each of the simulated stakeholders attributes to the policy intervention. For example, hotel guests with a sustainability orientation (D1A and D2A) might value highly the use of water saving devices in three-star hotels; while non-local three star hotel owners (S2+3*) who would be forced to install them, most likely will not. The point of σ_{acc} is to capture the degree to which the policy intervention harmonises the motivations and needs of all the simulated stakeholders. In this case non-local owners, the parameter "average water consumption per room per year in 3-Star hotels" decreases and this decrement affects indicators associated with water sustainability. However, while this adoption of water-saving devices has a positive impact in the customer satisfaction parameters of D1A and D2A stakeholders, it has a negative impact in the return on investment parameters of the owners of 3-star hotels (S1+3* and S2+3*) as well. And these changes might be reflected in their respective effectiveness and adequacy aggregation functions (calculated with respect to their own value interpretations). Hence 3-star hotel owners and sustainability oriented tourists would be more or less inclined to accept a policy intervention that includes the policy instrument "install water-saving devices".

6 Exploring policy interventions in a value-driven sandbox

The main purpose of a sandbox —in any policy domain— is to support a systematic exploration of alternative policy interventions in order to, eventually, choose one to deploy.

This exploration can be achieved along the lines suggested in Fig. 1 but testing potential interventions —eventually fine-tuning the model— needs to be carried out in a systematic way. Namely, one starts with a good-enough model and the starting conditions defined in one scenario. Then one feeds the model with a tentative policy intervention (instruments and evaluation assumptions) end assess how successful that intervention is according to the assessment assumptions. One would perform enough test runs to decide how appropriate that particular intervention is, and then modify the intervention in order to compare it with alternative interventions for the same scenario.

Heuristics for choosing alternative interventions are sketched in Fig.7. The gist is to start with a politically viable consensual intervention Π, test it and evaluate how good it is (core cycle). Then improve Π experimenting with the most likely improvements and stop when, according to Σ, good enough interventions are found. Without going into details, the secondary cycle ("**- - -**") for updating an intervention Π' is to first

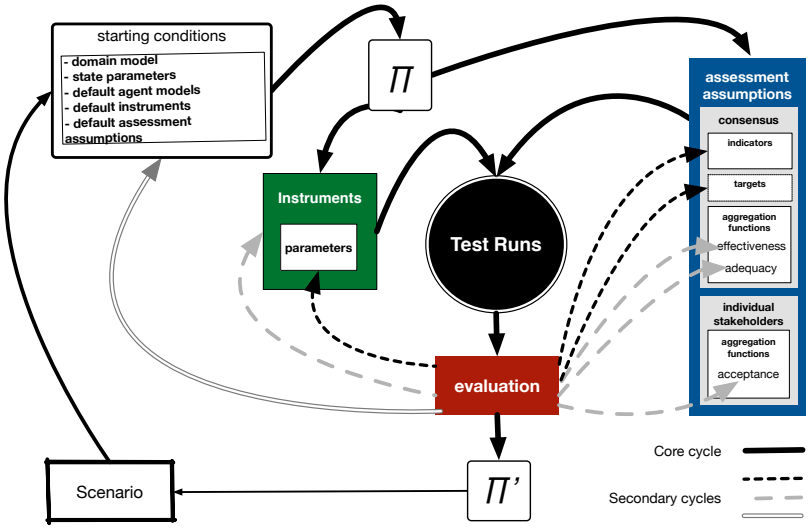

Fig. 7: Simulation flow: The selection of a policy intervention results from a simulation process that starts with some default conditions and triggered with a candidate policy intervention (Π) whose results are evaluated after a number of test runs (core cycle). The original components of the interventions (instruments and assessment assumptions, default assumptions and even the underlying models) are systematically modified (successive secondary cycles) in order to end up with an intervention Π' that is deemed satisfactory for a given scenario.

touch the instrument parameters, then the targets scores and then the indicators used in the effectiveness assessment function, go back and repeat this primary cycle with a different set of instruments. If this is not satisfactory, then modify successively the effectiveness, adequacy and acceptance functions. Once this cycle is settled with a good enough intervention Π', test alternative scenarios.

It is not unusual that this exploration reveals inadequacies of the simulated agent models, or the institutional and physical models. In this case, the exploration cycle is restarted with a parsimonious improvement of these models.

It is beyond the scope of this paper to discuss how one can develop, fine-tune and validate a model built along the criteria mentioned in the previous sections. Nevertheless we can point out that such tasks follow an analogous process that goes parsimoniously from Φ to Ψ, to M, to Σ (see [15]).

7 Closing remarks.

1. It goes without saying that sandboxes are only one of several tools to support policy design, assessment and negotiation. Not unlike such other tools, sandboxes require appropriate field-work and should be surrounded by use-practices that secure the involve-

ment of stakeholders in the process of design and approval a policy proposal. They also require adroitness and experience in order to bolster political commitment.

2. The distinctive advantage of sandboxes for policy design is grounded on their two main uses: as a tool for exploration of potential policies —specially unconventional ones— and as an evidence-based tool for negotiation among stakeholders.

3. The generic advantages of bringing agents as an essential part of modelling are to (i) to capture the effects of individual decision-making of stakeholders into the dynamics of the simulated policy domain, (iv) encapsulate AI-enabled actions into simulated agents (mostly of classes 4 and 5 in Sec. 4).

4. The key purpose of modelling values into an agent-based sandbox is to explicitly address the differences among stakeholders motivations and interests. The advantage of values-driven modelling is twofold.

First, as a tool for systematic exploration of policy effects, incorporating values in agent-based simulation produces a more nuanced representation of the policy environment and its dynamics, thus contributing to more realistic, flexible and reliable evidence-based predictions.

Second, values-driven modelling facilitates consensus on the choice of policy instruments and the assessment of the result of their use. Therefore facilitating the acceptability of policy proposals.

5. These two advantages of values-driven sandboxes —more reliable forecasts and more informative dynamics— suggest an improved ulterior use of the sandbox as an artefact for the follow-up and updating of a deployed policy: if the forecasted states differ from actual observed states, one should presume that the sandbox design is flawed (hence, one should refine it) while also presume that it is accurate and question its implementation.

6. We used the example of a tourist destination to illustrate one way of designing a values-driven sandbox for sustainable tourism. This illustration suggests how it can be scaled up to regional and supra-regional policy making. Likewise its adaptation to sustainability policies in other sectors.

7. This example also suggests how one can use other AI technologies —like "smart" data-gathering and processing, the use of AI enabled monitoring and oversight, AI-based nudging— in a given policy domain to modulate the operation of sustainability-sensitive infrastructure and make relevant information available and in a timely fashion. Similarly, how these and other AI technologies —machine learning, multiagent systems, generative AI edge-computing— can enable policy instruments of the four types mentioned in Sec. 5.

8. This example also suggests how to use values-driven agent-based simulation helps analyse paradoxical and undesirable trends in tourism destinations, like the trajectories towards and away from saturation, or the interplay between local and external ownership and operation of infrastructure and services.

9. The scientific and technological task of developing values-driven sandboxes is also a paradigmatic instance of the so called "value-alignment in AI", namely the design of artificial intelligent systems whose autonomy is objectively in accordance with human values ([25,14,17]).

Acknowledgements

Authors wish to acknowledge the contributions and fruitful discussions with Mark d'Inverno, Julian Padget, Enric Plaza and Harko Verhagen. Research for his paper is supported by CSIC's (Bilateral Collaboration Initiative i-LINK-TEC) project DE-SAFIA2030 BILTC22005; EU (Horizon-EIC-2021-Pathfinderchallenges-01) Project VALAWAI 101070930; and the EU (NextGenerationEU/PRTR program) and the Spanish (MCIN/AEI-10.13039–501100011033 program) project VAE TED2021-131295B-C31.

References

1. Camilla Adelle and Sabine Weiland. Policy assessment: the state of the art. *Impact Assessment and Project Appraisal*, 30:25 – 33, 2012.
2. Lia ní Aodha and Bruce Edmonds. Some pitfalls to beware when applying models to issues of policy relevance. In Bruce Edmonds and Ruth Meyer, editors, *Simulating Social Complexity: A Handbook*, pages 801–822. Springer International Publishing, Cham, 2017.
3. Linda C. Botterill and Alan Fenna. *Interrogating Public Policy Theory*. Edward Elgar Publishing, Cheltenham, UK, 2019.
4. P. Cairney. *The Politics of Evidence-Based Policy Making*. Palgrave Macmillan UK, 2016.
5. B. Dente. *Understanding Policy Decisions*. SpringerBriefs in Applied Sciences and Technology. Springer International Publishing, 2013.
6. Bruce Edmonds, Christophe Le Page, Mike Bithell, Edmund Chattoe-Brown, Volker Grimm, Ruth Meyer, Cristina Montañola Sales, Paul Ormerod, Hilton Root, and Flaminio Squazzoni. Different modelling purposes. *Journal of Artificial Societies and Social Simulation*, 22(3):6, 2019.
7. European Comission. *Better Regulation Toolbox*, n.d. [Online. Retrieved 2019, March 20].
8. Iason Gabriel. Artificial intelligence, values, and alignment. *Minds and Machines*, 30(3):411–437, 2020.
9. Nigel Gilbert, Petra Ahrweiler, Pete Barbrook-Johnson, Kavin Preethi Narasimhan, and Helen Wilkinson. Computational modelling of public policy: Reflections on practice. *Journal of Artificial Societies and Social Simulation*, 21(1):14, 2018.
10. Nicholas M. Gotts, George A.K. van Voorn, J. Gareth Polhill, Eline de Jong, Bruce Edmonds, Gert Jan Hofstede, and Ruth Meyer. Agent-based modelling of socio-ecological systems: Models, projects and ontologies. *Ecological Complexity*, 2018.
11. David Mair, Laura Smillie, Giovanni La Placa, Florian Schwendinger, Milena Raykovska, Zsuzsanna Pasztor, and Rene Van Bavel. Understanding our political nature: how to put knowledge and reason at the heart of political decision-making. Technical report, Joint Research Centre, 2019.
12. Peter J. May. Policy design and implementation. In B.G. Peters and J. Pierre, editors, *The SAGE Handbook of Public Administration*, pages 279–291. SAGE Publications, 2nd edition, 2012.

13. Pablo Noriega and Enric Plaza. On Autonomy, Governance, and Values: An AGV Approach to Value Engineering. In Nardine Osman and Luc Steels, editors, *Value Engineering in Artificial Intelligence*, pages 165–179, Cham, 2024. Springer Nature Switzerland.
14. Pablo Noriega and Enric Plaza. The Use of Agent-based Simulation of Public Policy Design to Study the Value Alignment Problem. In Pompeu Casanovas, editor, *Artificial Intelligence Governance, Ethics and Law (AIGEL).*, CEUR Workshop Proceedings, pages 60–78. CEUR, In press.
15. Pablo Noriega, Harko Verhagen, Julian Padget, and Mark d'Inverno. Ethical online AI systems through conscientious design. *IEEE Internet Computing*, 25(06):58–64, nov 2021.
16. Pablo Noriega, Harko Verhagen, Julian Padget, and Mark d'Inverno. Design heuristics for ethical online institutions. In Nirav Ajmeri, Andreasa Morris Martin, and Bastin Tony Roy Savarimuthu, editors, *Coordination, Organizations, Institutions, Norms, and Ethics for Governance of Multi-Agent Systems XV*, pages 213–230, Cham, 2022. Springer International Publishing.
17. Pablo Noriega, Harko Verhagen, Julian Padget, and Mark d'Inverno. Addressing the value alignment problem through online institutions. In Nicoletta Fornara, editor, *Coordination, Organizations, Institutions, Norms, and Ethics for Governance of Multi-Agent Systems XVI*, page in press, Cham, in press. Springer International Publishing.
18. Douglas North. *Institutions, Institutional Change and Economic Performance*. Cambridge University Press, Cambridge, UK, 1991.
19. Nardine Osman and Mark d'Inverno. A computational framework of human values. In *AAMAS'24*, pages 1531–1539. IFAAMAS / ACM, 2024.
20. Nardine Osman and Luc Steels, editors. *Value Engineering in Artificial Intelligence - First International Workshop, VALE 2023, Krakow, Poland, September 30, 2023, Proceedings*, volume 14520 of *Lecture Notes in Computer Science*, Cham, 2024. Springer.
21. Antoni Perello-Moragues and Pablo Noriega. A playground for the value alignment problem. In Lourdes Martínez-Villaseñor, Ildar Batyrshin, and Antonio Marín-Hernández, editors, *Advances in Soft Computing. Mexican International Conference on Artificial Intelligence*, volume 11835 of *LNCS*, pages 414–429. Springer, 2019.
22. Antoni Perello-Moragues and Pablo Noriega. Using agent-based simulation to understand the role of values in policy-making. In *Advances in Social Simulation*, pages 355–369. Springer, 2020.
23. Antonio Rangel, Colin Camerer, and P. Read Montague. A framework for studying the neurobiology of value-based decision making. *Nature Reviews Neuroscience*, 9(7):545–556, 2008.
24. Stuart Russell. Of myths and moonshine. *EDGE*, 2014. Accessed: 2021-12-01.
25. Stuart Russell. Of Myths and Moonshine. A conversation with Jaron Lanier, 14-11-14. The Edge, November 2014. [Online] Retrieved 18 December 2023.
26. Carles Sierra, Nardine Osman, Pablo Noriega, Jordi Sabater-Mir, and Antoni Perelló. Value alignment: a formal approach. *CoRR*, abs/2110.09240, 2021.
27. Herbert A. Simon. *The Sciences of the Artificial*. MIT Press, third edition, 1996.
28. Herbert A Simon. Fact and Value in Decision-making. In *Administrative Behavior: A study of decision-making processes in administrative organization*. The Free Press, 4th edition, 1997.
29. Christopher M. Weible and Paul A. Sabatier. *Theories of the Policy Process*. Routledge, 2018.