AI & SOCIETY
https://doi.org/10.1007/s00146-022-01518-8

OPEN FORUM

Maximizing team synergy in AI-related interdisciplinary groups: an interdisciplinary-by-design iterative methodology

Piercosma Bisconti[1] · Davide Orsitto[1] · Federica Fedorczyk[1] · Fabio Brau[2,4] · Marianna Capasso[3] · Lorenzo De Marinis[2] · Hüseyin Eken[3,4] · Federica Merenda[1] · Mirko Forti[1] · Marco Pacini[5,6] · Claudia Schettini[1,4]

Received: 14 June 2021 / Accepted: 1 June 2022

Abstract
In this paper, we propose a methodology to maximize the benefits of interdisciplinary cooperation in AI research groups. Firstly, we build the case for the importance of interdisciplinarity in research groups as the best means to tackle the social implications brought about by AI systems, against the backdrop of the EU Commission proposal for an *Artificial Intelligence Act*. As we are an interdisciplinary group, we address the multi-faceted implications of the mass-scale diffusion of AI-driven technologies. The result of our exercise lead us to postulate the necessity of a behavioural theory that standardizes the interaction process of interdisciplinary groups. In light of this, we conduct a review of the existing approaches to interdisciplinary research on AI appliances, leading to the development of methodologies like *ethics-by-design* and *value-sensitive design*, evaluating their strengths and weaknesses. We then put forth an iterative process theory hinging on a narrative approach consisting of four phases: (i) definition of the hypothesis space, (ii) building-up of a common lexicon, (iii) scenario-building, (iv) interdisciplinary self-assessment. Finally, we identify the most relevant fields of application for such a methodology and discuss possible case studies.

Keywords Interdisciplinarity · Artificial intelligence · AI regulation · Ethics-by-design · Design methodology

✉ Piercosma Bisconti
piercosma.biscontilucidi@santannapisa.it

Davide Orsitto
davide.orsitto@santannapisa.it

Federica Fedorczyk
federica.fedorczyk@santannapisa.it

Fabio Brau
fabio.brau@santannapisa.it

Marianna Capasso
marianna.capasso@santannapisa.it

Lorenzo De Marinis
lorenzo.demarinis@santannapisa.it

Hüseyin Eken
huseyin.eken@santannapisa.it

Federica Merenda
federica.merenda@santannapisa.it

Mirko Forti
mirko.forti@santannapisa.it

Marco Pacini
mpacini@fbk.eu

Claudia Schettini
claudia.schettini@santannapisa.it

[1] Dirpolis Institute, Sant'Anna School of Advanced Studies, Piazza dei Martiri della Libertà 33, Pisa, Italy

[2] Tecip Institute, Sant'Anna School of Advanced Studies, Piazza dei Martiri della Libertà 33, Pisa, Italy

[3] Biorobotics Institute, Sant'Anna School of Advanced Studies, Viale Rinaldo Piaggio, 34, 56025 Pontedera, Italy

[4] Department of Excellence in Robotics and AI, Scuola Superiore Sant'Anna, Piazza Martiri Della Libertà, 33, 56127 Pisa, PI, Italy

[5] Fondazione Bruno Kessler (FBK), Via Sommarive, 18, 38123 Povo, TN, Italy

[6] University of Trento, Trento, Italy

Published online: 28 June 2022

1 The problem of regulating AI

Artificial intelligence (AI) is nowadays at the centre of the public debate of the International Community and within the most developed states. In fact, in these very months, European policymakers are grappling with the challenging task of regulating the innovative power of such an array of technologies so that they can be employed in the multi-faceted reality of our everyday life. This objective is pursued while, at the same time, respecting the ethical principles that are diffused by already existing norms, processes, practices that our societies are based upon.

In October 2020, the European Parliament (EP) adopted a set of resolutions (EU Parliament, 2020/2012–2014-2015 INL) to provide the Commission with recommendations to oversee the development of the upcoming European legislation on AI. On April 21st, 2021 the Commission published a regulation proposal for an "Artificial Intelligence Act" (EU Commission 2021).

It is relevant to state that the Commission did not accept all EP recommendations. In particular, one of the EP resolutions dealt with a *Framework of ethical aspects of artificial intelligence, robotics and related technologies* (EU Parliament 2020/2012). The EP document laid out key principles regulating the development and use of AI systems in different contexts: transparency, explainability, fairness, accountability and responsibility 2019). The resolution mentioned such contexts explicitly while recalling the work of the Independent High-Level Expert Group on Artificial Intelligence (AI HLEG), established by the Commission, which elaborated the Assessment List for Trustworthy Artificial Intelligence (ALTAI, HLEG 2020). It furthermore endorsed a self-assessment procedure based on the ALTAI methodology as a tool to ensure the respect of such principles.

As in the regulation proposal, the Commission identifies four categories of AI, basing on a risk-assessment, with respect to "high-risk" AI, the Commission deemed self-assessment the most appropriate tool to enact the regulation on AI within the framework of a more comprehensive system that hinged on the issuance of a *European Certificate of Ethical Compliance* (EU Commission, 2021 Chapter 5). The new system tasks the national authorities with monitoring the compliance to the regulation, and entrusts the executive and coordinating functions to a future centralized *European Artificial Intelligence Board* (EU Commission 2021, title VI). However, the reference to the ALTAI guidelines has been removed from the proposal's final text, leaving the subjects to perform the self-assessment procedure without clear and specific indications and guidance.

As a result, all the different actors to be involved in such a process will have to develop tools and procedures to perform their duties under the future regulation, which makes the cooperation between subjects having different expertise on AI and thus the creation of Interdisciplinary Research and Working Groups (IRWGs) spanning different perspectives (e.g., legal, ethical, economic, social, engineering) on AI a much needed asset in the near future.

Even though IRWGs will soon prove vital to guide the implementation process of the AI regulation in Europe, as of today there is no effective organizational methodology that allows the Interdisciplinary group to maximize the output of their activity. IRWGs are indeed nowadays mainly left to spontaneous interaction. This raises the necessity to develop a methodology that builds the process that IRWGs can employ to maximize the output of their activity. The objective of this paper is to turn the hurdles that a group with competences spanning various disciplines has in its iterative phase, to develop a methodological mechanism to work on AI that could integrate a multidisciplinary approach to AI design. In the next sections, we conduct a literature review of the existing approaches to AI design, further strengthening the case to introduce a process theory specific to multidisciplinary groups. Such a theory sets forth pre-arranged modes of interaction that best reduce the lack of mutual understanding and other shortcomings of IRWGs. We then develop the guidelines of an interdisciplinary approach by putting forth an iterative methodology based on the narrative approach in four phases: (i) definition of the hypothesis space, (ii) building up of a common lexicon, (iii) scenario building, (iv) and interdisciplinary self-assessment. Furthermore, we show how our methodology can improve the design and the assessment process of AI and AI-related technologies in practice, by discussing the cases of AI algorithms for Judicial Trials and in the field of Social Robotics.

2 A review on existing interdisciplinary approaches to AI-design

The issue of employing AI systems in societal environments requires an interdisciplinary approach to balance the technical issues and forecast biases, paying attention to the weight that these measures have on policy decisions. Before proposing a solution, we discuss some of the existing methodologies that attempt to endogenize the social impact variable in the phase of technological design. Among the different approaches, we consider the philosophical perspective of *Ethics by design,* since it is becoming a progressively important approach to AI development. At the beginning of the history of AI development, producers were driven to design technological tools for practical purposes. After the "Empirical Turn" of the 1990's, they started focusing also on the ethical aspects of design.

Such an approach considers designers to be moral agents looking at the ethical dimension of technological design as the result of both individual and collective action. This field of study emerged during the mid-1990s, through the work of scholars such as Alain Findelli and Carl Mitcham who addressed the philosophical issues regarding the ethics of design. In his *"Ethics into Design"*, Mitcham argues that the two traditional visions of design developed in the twentieth century— *design as art* and *design as a scientific and logical process* —"must be complemented by the introduction of ethics into design, in order to contribute to the development of a genuinely comprehensive philosophy of design" (Mitcham 1995). *Ethics by Design* essentially refers to an organizational approach envisioning a responsible use of technology. This approach is related to classical issues in ethical philosophy and law which are transposed into the realm of intelligent machines. To fully benefit from the potential of AI, one needs to ensure that such technologies, which are nowadays acquiring more and more independence, be aligned with societal moral values and ethical principles to behave in a human-friendly way.

In a nutshell, *Ethics by Design* concerns "the methods, algorithms and tools needed to endow autonomous agents with the capability to reason about the ethical aspects of their decisions, and the methods, tools and formalisms to guarantee that an agent's behaviour remains within given moral bounds" (Dignum et al. 2018). In the *Ethics by Design* framework, two issues are central: research methodologies and design processes that deal with the issue of developing technologies that are compliant to human values such as accountability, responsibility and transparency; and the analysis of the ethical and social dimensions that are embedded in technological decision-making processes. The former issues pose a specific call for interdisciplinary communities to define, develop and regulate the profound impacts that technologies such AI may have on society as a whole. As such, beyond the reflection on programming machines that behave according to some ethical principles and norms, *Ethics by Design* primarily aims to investigate possible responsible methodologies that require the cooperation of different stakeholders.

The issue of responsibility also concerns the use of personal data and the transparency of AI services: the *Value Sensitive Design* approach provides concrete design guidelines on how to build "ethics" into the design and development of technological devices. *Value-sensitive design (VSD)* is defined indeed as "a theoretically grounded approach to the design of technology that accounts for human values in a principled and comprehensive manner throughout the design process" (Friedman and Hendry 2019). The main aim of this approach is to proactively include the consideration of moral and social values in the design and implementation of technological systems.

VSD was originally developed in the domain of information and communication technologies (ICT) but it is now applied in different domains so that scholars prefer to talk about *Design for Values (D4V)* (Van den Hoven, Vermaas, and Van de Poel 2015). Specifically, its methodology is based on three iterative investigations: conceptual, empirical and technical. The conceptual investigation involves the identification of direct and indirect stakeholders, values and value trade-offs that are or will be affected by the technological systems at stake. Conversely, the empirical investigation evaluates "understandings, contexts, and experiences" (Friedman et al. 2020) of stakeholders and translates them into design requirements. Finally, the technical investigation deals with the architecture and features of technological systems to identify how those latter can implement and support the values elicited in the first two investigations.

Therefore, the central tenet of VSD is that values should be made explicit and transparent and operationalized as non-functional design criteria ab initio and throughout the design-in-progress of a system, allowing the designers to modify their interventions continuously (Christen et al. 2020).

VSD has sparked some controversy. Criticisms are related to the fact that VSD cannot adequately distinguish between stakeholders' preferences and moral values (Manders-Huits 2011) and that it lacks a moral commitment to any ethical theory as such (Reijers and Gordijn 2019). Approaches such as *Care-Centred Value Sensitive Design* (CCVSD) have been developed as a response to such criticisms with the aim to normatively ground VSD on an explicit reference to care ethics (Van Wynsberghe 2013).

Another challenge for VSD is brought up by the ethical design of AI-driven systems. As a matter of fact, scholars have recently argued that since such systems are autonomous, adaptive and interactive, they provide unique value-embedding opportunities that were never intended by the initial system designers. Thus, AI systems should undertake a more continuous monitoring and redesign, in comparison to traditional socio-technical systems (van de Poel 2020).

Following the premises provided Ethics by Design and VSD, the involvement of diverse developer groups in AI systems highlights the potential of interdisciplinary collaborations to have a better overview of the problems from different perspectives; propose solutions to overcome such problems with underlying ethical, philosophical, legal, and economic requirements integrated into the design process; and subsequently conduct more precise feasibility assessments of the possible solutions. To achieve these goals, several researchers have articulated why and how the involved parties with different backgrounds should work in collaboration. Dwivedi et al. (2019) presented a collective insight from several expert contributors with backgrounds such as business and management; arts, humanities, and law; science

and technology; government and public sectors. The contributors have emphasized the opportunities, assessed the impacts, and identified the current challenges and research directions of their fields, in response to the rapid emergence of AI. This work constructs the first step towards interdisciplinarity and presents a good starting point for IRWGs to identify strengths and weaknesses of AI applications reflected on different domains. Another work focusing on AI explainability by Beaudouin et al. (2020) introduced a more collaborative work of technical, economical, and legal teams. After defining the problems from their perspectives, the teams defined the "right" level of case-specific explainability by conducting a cost–benefit analysis, in which explainability is deemed plausible if the societal benefits exceed the associated costs. The interdisciplinary approach, particularly on defining the costs, allowed for a more accurate assessment since the associated costs arise from both technical, economical, and legal aspects.

3 The case for an interdisciplinary approach to AI design and development

While the Ethics-by-Design approach and VSD try to develop methodologies r to embed the concepts of ethics and responsible design into the *production* processes of General-Purpose Technologies including AI, a wide need for interdisciplinarity has been raised inside the *academic* community and the *corporate* world.

The methodologies presented in the previous section do not provide a formalized process to ensure that research groups themselves are in the condition to face the multifaceted issues in an interdisciplinary way. In fact, the assessment of the value and social impact, as well as the ethical framework of a technology is a complex matter that requires broad competences spanning different fields of knowledge. It can be often seen how disciplines such as sociology, psychology, philosophy, and engineering approach the challenges involving technology through their different backgrounds: this means that different understandings and framings of the issues at hand are analysed. The renowned challenge concerning the design of General-Purpose Technologies is precisely the seemingly unbridgeable distance between technical and social sciences. Such a distance often ends up in an *ex-post* assessment of the ethical and value-related compliance of innovative technologies.

The main shortcoming of the above-mentioned approach is that engineers might see the sociological and ethical assessment as a formal hurdle to technology design. On the other hand, social scientists too often do not have a clear grasp of the technology they are evaluating. A good example of this demarcation issue is the discussion in Social Robotics concerning the "deception objection", namely the possibility that Social Robots will deceive users on their mechanical nature (Sharkey and Sharkey 2012). The discussion on the "deception objection" has shown progress in the last years, and it has been characterized by a deep divide between the technical and sociological fields, where sociologists and philosophers did not understand immediately that today's machines would never be able to deceive a user about their real nature. Only after years the focus moved to the possibility of self-deception, far more likely (Coeckelbergh 2018). The discussion could have been focused from the beginning on more urgent issues, or on a more precise framing of the "deception objection". Moreover, interdisciplinarity needs to remain a key principle to follow to effectively implement the EU proposal of AI regulation (EU Commission 2021). For all these reasons, it is urgent to develop a sound interdisciplinary process to overcome these limitations.

In this regard, it should be taken into consideration that AI can also be used in contexts where fundamental rights are at stake, e.g. in criminal trials to determine the likelihood of recidivism. As we will address in chapter 5, although in Europe the use of similar technologies in criminal matters still remains an unlikely event/hypothesis, it is important to consider this possibility, since it cannot be excluded a priori.

The most relevant objection raised against the introduction of similar tools in the European framework came from European data protection law: according to Article 22 of GDPR, data owners have the right not to be subjected to a solely automated decision, especially when the decision produces legal effects concerning the individuals or similarly significantly affects them.

As a consequence, coherently with the European «human-centric» approach to AI, the final outcome of the decision process should derive from human agency, and automated decision systems can only play a secondary role of assistance in elaborating relevant information and dealing with preliminary results. However, who can ensure that the judge is granted access to unbiased data to decide his/her ruling? Once the algorithm gives way to an outcome, it is unlikely that the judge shoulders the responsibility of reversing her decision. The algorithm is undoubtedly an instrument of strong psychological pressure for the judge, who would be induced not to deviate from its results. The key issue on which the analysis hinges no longer only concerns the likelihood that the AI assistance output for the criminal trial is biased, but rises to the general quandary of how to create a process that is able to correctly govern AI in the trial. In other words, the debate must address *how* judicial systems will deal with these technological developments, without becoming their victims. Such conclusions pave the way to the need to define a methodology that regulates the design process of AI systems by multidisciplinary research and working groups. Our aim is not to replace VSD and ethics-by-design but to provide these methodologies with

a second-level methodology that allows us to manage the interdisciplinarity of the research process itself.

4 An iterative methodology for interdisciplinary research groups

In this chapter we design a methodology to ensure an *interdisciplinary-by-design approach* to IRWGs. The process of this methodology is divided into five steps.

4.1 A narrative approach

The narrative paradigm as an assessment framework refers to the use of stories that exhibit coherence and fidelity in explaining human communication and interpreting or assessing public issues and group relationships (Fisher 2021). The use of 'narrativity' has been conceptualized by several different theories, especially in cases concerning an interplay of technologies with societal or political issues. According to such perspectives, technologies embody what scholars have called a 'narrative capacity, in that they actively shape narratives and the understanding thereof by introducing new events and characters in the public and cultural sphere of actors (Coeckelbergh and Reijers 2016; Reijers and Coeckelbergh 2020).

In this respect, the adoption of any narrative approach at the first stages of the process may be a propaedeutic method to understand and define the developments of specific AI-driven technologies in certain contexts. More specifically, the employment thereof may also strengthen the cooperation of multiple stakeholders, with no technical competences, to critically outline and discuss the ongoing impacts of their work on AI. The use of a narrative working process is especially relevant. An instance of it may come in the shape of storytelling that represents different perspectives, expectations, future images and possible outcomes. This can improve the stakeholder engagement process as well as helping stakeholders to frame a participatory process (Quick 2018). Therefore, we adopt a narrative approach since it better fits the multiplicity of standpoints and competences that a IRWG might represent. The methodology we propose is divided in three stages where the narrative sharing aims to reach two goals:

1) Let the different standpoints and understandings of the functioning and impacts of AI systems emerge.
2) Provide a common and shared background between the researchers, before the design phase.

In practice, the three stages will take place during a focus group.

4.2 Definition of the hypothesis space

One of the main goals of the proposed methodology is to create value from the heterogeneity of backgrounds and perspectives represented by the different members of the multi-stakeholder expert groups. An implicit goal to meet during the process is therefore the transformation of such differences from a possible challenge into an epistemic resource. In fact, these differences often constitute an obstacle to the development of socio-technical systems that are satisfying for all the actors involved. In light of such goal, the first step is the setup of an interactive dialogue among the IRWG, aiming to reach a common definition of the "hypothesis space". The term is drawn from machine learning studies (Blockeel 2011), which defines as "hypothesis space" the set of possible solutions among which the training algorithm will choose the most appropriate model to execute the required task in respect to the available data.

The definition of which solutions should in the researcher's opinion belong to the hypothesis space is to some extent an epistemological question, in that it significantly depends on the reference system of their disciplinary perspective. The realm of the *possible* solutions to a certain problem depends on the constraints that are imposed over such set, e.g., if the goal is to find the most appropriate AI system to perform a certain task, for a data scientist the hypothesis space will correspond to the set of existing technologies which are able to perform that specific task, whereas for an expert in ethics the hypothesis space will correspond to the set of technologies which while performing that task do not violate the chosen ethical values, for a legal expert corresponds to the set of technologies which respect legal constraints and so on.

In order for the group to identify solutions that are possible and thus appropriate from the perspectives of all the disciplinary fields included, the hypothesis space of each field must be traced so that a common hypothesis space can emerge from the intersection of all the sectorial hypothesis space considered.

4.3 Building-up of a common lexicon

In this phase, the different stakeholder communities engage in a general process of disambiguation on the terminology and conceptualization of their work on AI. Such a process should go beyond the practical design challenges related to a specific AI-driven technology, to include a prior assessment of different stakeholder communities involved in AI-related research with their own peculiar epistemic frameworks and norms. Indeed, such disambiguation may be a means to understanding how such communities adopt lexicon definitions of AI-related research that may either overlap or diverge. For instance, according to Preece et al. (2018), it should be possible to reach 'composite' and

'layered' explanations, whose function is to incorporate or unpack—when required—the information needed to satisfy multiple stakeholder communities, notwithstanding their different concerns and their different conceptions (i.e., on transparency, explicability, and so on) (Preece et al. 2018). The goal of this phase is to produce a lexicon map of the different meanings attached to the key concepts involved in the design/implementation of the AI system.

4.4 Scenario-building

The phase of scenario building aims to narrow down the construction of a common understanding of the usage, benefits, and risks of a given technology between the researchers and the involved stakeholders. The second phase regarding the setting up of the hypothesis space envisions the possible solution that a given technology supplies, considering the multiplicity of reference systems of the various disciplines involved. A common lexicon paves the way to the awareness of the different meanings and standpoints toward the same word or concept. Subsequently, the co-creation of a reality-check process is required: the reached common understanding should produce the co-creation of a scenario-building where people from different backgrounds can test whether the shared knowledge produces also a common and homogeneous understanding of the issues involving the given AI system. Therefore, the scenario building has three features:

- The interdisciplinary group faces a precise and concrete application of the technology in a real use-case situation. This helps to avoid that the excessive abstraction may cause misunderstandings or agreements between the group members that are based on diverging interpretation and constructive ambiguities.
- The group will first decide what are the parameters on which an interdisciplinary common understanding must be achieved to create the scenario. This involves the employment of the General-Purpose Technology, its risks of provoking harm and the expected benefits. Obviously, the laying out of all the possible issues pertaining to an interdisciplinary group dealing with the design, regulation, and marketing of a technology is beyond the scope of the paper. We wilfully streamline our analysis by postulating that the three concepts of usage, risks, benefits are to be taken into account.
- Each member will set up the fundamental Key Performance Indicators (KPIs) that, in her perspective, the given technology should provide in order to be effective. These KPIs will refer to the parameters. This particular step comes at the very end of the interdisciplinary assessment to ensure that the KPIs proposed by each member will be sensitive to the different understanding shared in the previous phase. The goal of this phase is that each member provides her standpoint toward the scenario, formalizing the KPIs.

4.5 Interdisciplinary self-assessment

In this phase, the degree of agreement on the concepts, meanings and outcomes of every parameter and KPI must be registered. This involves an interdisciplinary self-assessment, based on the coherence between the KPIs proposed by each member. From the results of this self-assessment two outcomes arise:

- The degree of mutual understanding and meaning co-construction inside the interdisciplinary group is considered sufficient. The interdisciplinary setting of the group is concluded.
- There is still a high degree of ambiguous or obscure meanings, non-homogenous evaluation of parameters or any other intersubjective misunderstanding. At this stage, the process should start over from the definition of the hypothesis space.

The goal of this stage is to assess the coherence and the homogeneity of the KPIs envisioned by each member. On the basis of this output, the interdisciplinary process might end or start from the beginning.

We have proposed a methodology to allow interdisciplinary groups to effectively deal with the complexity of innovative technologies as AI systems. As suggested above, one of the main open issues of technological innovation is the multiplicity of implications (technological, social, political, psychological, legal) that it might bring about. The simple fact that a group of researchers is interdisciplinary does not ensure that the design of a given technology will be tackled in an interdisciplinary way. This often affects trustworthiness, legal compliance and disregards social implications. On the other hand, the discussion between non-technical scholars, civil society and in general non-technical stakeholders often lacks a rigorous understanding of AI. To bridge the gap between these two requires a methodology that does not deal with the design in itself, but with the process that brings an interdisciplinary group to design the technology.

We proposed to spread such an effort into the above-mentioned phases, with a final self-assessment procedure. Firstly, the definition of the hypothesis space will enquire how the understanding of what is a "solution" might differ in an interdisciplinary group. Moreover, the definition of a common lexicon will allow us to co-create and compare how different epistemic communities construct the terms and the concepts, to understand how they diverge or overlap. This disambiguation is necessary to avoid misunderstandings. Finally, the narrative approach will be applied to a

scenario: a precise application of the technology. After the definition of parameters on which the technology is evaluated, each member will set the KPIs of the technology in that given use case. This will narrow down the expectations of each member and the different understanding of the given technology.

5 Fields of application

The application of an interdisciplinary methodology, as the one described, would be of great help for groups of researchers dealing with technology with complex socio-technical impact as Artificial Intelligence or AI-related technologies. The narrative approach that we envisioned would be helpful mainly to ensure that a common understanding of the group's different standpoints is reached, and to avoid that interdisciplinarity becomes only a buzzword.

Indeed, in general, AI-driven risk assessment algorithms replicate the human process: learning information, communicating results and making a decision. However, the concrete experience of certain algorithms revealed non-negligible problems, with potential negative consequences on fundamental values and rights. Recently, the employment of AI for risk-assessment tools has been extensively researched in (Green and Chen 2021). In particular, the authors focus on the impact of the model evaluation process, observing that because many policy decisions entail balancing risk reduction with competing goals, increasing prediction accuracy may not necessarily enhance decision quality.

The following is a series of examples where the proposed interdisciplinary approach might enhance the effectiveness and prevent the shortcomings of AI systems.

Thinking, for instance, of the increasingly widespread use of artificial intelligence and predictive algorithms in criminal trials worldwide: numerous American States are using software such as COMPAS to determine whether suspects should be incarcerated or not before trial, as well as determining the likelihood of their recidivism; likewise, the United Kingdom is using software comparable to COMPAS, named Harm Assessment Risk Tool (henceforth HART), which establishes if a suspect should be held in pre-trial detention or not (Ebers and Gamito 2020).

In this regard, COMPAS has been harshly criticized, due to racial and gender bias introduced during the design phase (ProPublica 2016; Flores, Bechtel, and Lowenkamp 2016; Rudin 2019; Wachter et al. 2017). Moreover, being COMPAS developed using non-explainable AI models and so furnished to judicial bodies as a black-box tool, these analyses were not able, by-design, to exploit the full knowledge of the model architecture and the internal parameters. Through the proposed methodology, a heterogeneous group of experts would have discovered the necessity to use accessible and interpretable models within the scenario building and self-assessment phase.

Another indicative example is HART, the Risk Assessment Tool used in the UK to predict the suspect's probability of having committed the crime and the probability of his or her recidivism. HART employs an interpretable model and uses 34 different inferential variables, some of which code for the person's gender and two forms of associated postcode. Postcodes and other geographical data may negatively affect the model due to variance in the presence of variables that affect the crime rate.

As a matter of fact, a high concentration of police forces in a particular area may inflate the Algorithm's calculations regarding the probability of the suspect committing the crime and his or her recidivism. This happens because a strong presence of law enforcement significantly increases the discovery rate of crimes. Such a phenomenon leads to a self-fulfilling prophecy, as the increase in the discovery of crimes enhances the crime rate, that in turn increases the concentration of police forces in that area, triggering an undesired loop in the justice mechanism. With an interdisciplinary approach by design, the chance to discover the "self-fulfilling prophecy" issues would have been more likely, since this is a well-known problem in sociology (Henshel 1982).

The employment of similar AI tools in the United States and UK prefigures their future spread also in Europe. In this regard, the AI Act already recognized the significant degree of power imbalance in the implementation of similar tools, since they may lead to arrest or deprivation of personal liberty and therefore assess fundamental rights guaranteed by the EU Charter of fundamental rights. Precisely for this reason, the Proposal classified as *high risk* the AI systems intended to be used in the law enforcement context (and, in particular, the one used by law enforcement authorities for individual risk assessment in the course of detention, investigation and prosecution of criminal offences), stressing the necessity of their transparency and explainability.

Being high risk systems, they should be provided with a risk management system that shall consist of a continuous iterative process run throughout the entire life-cycle of the AI system and the reduction of risks should be implemented from early in the first phase of design and development. Therefore, the proposed interdisciplinary methodology may play a crucial role in the process, trying to correct the flaws that have emerged from the analysis of the tools used overseas.

Indeed, the internal functioning of the aforementioned algorithms is mostly obscure, with the consequence that two defendants accused of the same crime may receive different treatments based on other information, unknown to them, with no possibility to challenge the results.

Furthermore, discriminatory effects are inherent in the use of these tools and often detached from them, since their output score depends on the parameters of the model, which depends on both available data and human choices.

Usually, the information needed to build the model is, in most of the cases: currently pending charges, prior arrest history, involvement in previous pre-trials, failure, residential stability, employment status, community ties, and substance abuse; according to the information that humans decide to include, the result will be different and could include or not undesired discriminatory effects. The problem that we highlight here is the process that brings some type of information to be included in the model, or not. From an interdisciplinary point of view, these models shall not only be auditable and explainable, but to be by-design approached from an interdisciplinary perspective. Different expertise may in fact envision the social implications of selecting given information. The process methodology we described in the previous section well adapts to this case: the phase of defining the hypothesis space should narrow down the actual aim and solutions that the given technology is supposed to provide (e.g. help—not substitute—judges in a trail). The scenario building phase allows the team to preventively tackle the issues of fairness, inclusivity that COMPASS and HART raised.

Hence, interdisciplinary groups are the best candidate that, under a proper methodology, to correctly select the fair information to use.

Another field of application for an interdisciplinary iterative methodology where AI plays a crucial role is surely interaction design for social robots. As discussed above, social robotics has been an interesting field where the need for an interdisciplinary approach emerged quickly and abruptly. On the side of engineering issues, social robotics immediately raised important design issues regarding the interaction. Engineers quickly realized the complexity of interactional patterns and evaluation scales for interactions. This complexity needed to be addressed simultaneously by different disciplines as in the case of robots approaching a human. In this simple action are involved not only engineering evaluations: proxemics (Mumm and Mutlu 2011), eye contact (Mutlu et al. 2009), gestures and facial expressions (Chumkamon et al. 2016), and obviously verbal interactions required psychology, cultural anthropology, sociology, philosophy, linguistics and many other disciplines in order to make the interaction effective—think for example to the complexity of the uncanny valley effect. Against this backdrop, we agree with Seibt et al. (2018) in calling for an integrative approach to social robotics. But for social robotics the issue of interdisciplinarity regards not only the design procedures—as the production of psychometric scales as ALMERE—but first and foremost the assessment of the possible effects of these technologies in society. As we discussed above, the lack of an interdisciplinary understanding of the complexity faced with social robotics brought scholars in discussing lateral issues, as deception, that are today purely fictional. On the other hand, interesting discussions emerge when interdisciplinarity is adopted: the reflection for example on robots' genderedness (Steinhaeusser et al. 2021) and the social implication of giving gender to robots raised the problem of inclusivity of these technologies.

The methodology proposed in this paper, if applied rigorously, would enable an *interdisciplinary-by-design* approach to social robotics. The definition of a hypothesis space and a common lexicon might have faster brought the discussion on deception to the one of self-deception, pointed out by Turkle years ago (Turkle 2007), but only recently become a central issue. In the scenario building phase, the presence of experts in gender studies could tackle from the design phase the implicit linkage between physical aspects of robots and their gendered role.

In this section we showed how an interdisciplinary process methodology could tackle since the design phase issues raised by AI algorithms and AI-related technologies. We claim that the application of an interdisciplinary process methodology could enhance the social benefits of AI (and related) technologies and allow us to tackle in advance the possible shortcomings. In light of the proposal for regulation from the EU Commission, we claim that risk-management procedures for AI algorithms should by-design require an interdisciplinary approach, in order to enhance the human-centeredness of technological innovation. While this work aimed at presenting the methodology on its concepts and framework, in future works we aim to report an applied case of this approach in a lab-based scenario.

6 Conclusions

In this paper we discussed the relevance of interdisciplinarity for the implementation of AI systems with multi-faceted social implications. We pointed out that, to correctly forecast social issues pertaining to the use of AI, it is necessary that ICRW interactions are proceduralized through a standard process theory. Conversely, the different conceptions of an AI-driven technology among researchers spanning different fields hinders the effectiveness of interdisciplinary approaches. We have argued how the implementation of the EU regulation on AI requires an effective interdisciplinary approach to tackle social implications engendered by the progressive integration of AI driven technologies in the EU Single Market. We discussed two fields where an interdisciplinary approach shall be required in order to effectively tackle the shortcomings of innovative technologies, such as AI systems for trials and Social robotics. Such shortcomings can be addressed only if different competences, building a common understanding of the topic, are brought together.

From this emerges the need to develop a methodology to formalize an interdisciplinary process aimed at increasing the synergies of ICWR research groups dealing with AI systems with urgent social implications. Without a process theory that enables effective group communication and understanding, interdisciplinarity is not sufficient to ensure that the solutions and implementations of AI systems are addressed in the best manner.

We have thus laid out an interdisciplinary process based on the narrative approach. In the first stage the narrative approach is conducted to facilitate the communication of different understanding on the same issue. The second stage deals with how the *hypothesis space* is defined. Such a stage is necessary to correctly understand what the possible solutions and applications of the AI systems are. In the third stage a common lexicon is set up. This ensures that interdisciplinary groups avoid falling short of their objective due to ambiguous or obscure meaning. Reaching a common understanding of the various meanings, rooted in the different disciplines, is a key requirement in order to have effective interdisciplinarity. A third necessary step deals with narrowing down the possible applications and implications of the AI system in a real-world scenario. The narrative that each member of the group offers on the one side highlights the different conception of the social implications spurring from the employment of AI. On the other side, it will show if the group of researchers reached an interdisciplinary common ground to analyse the implications of AI systems. The last stage of the process, the group will follow up with a self-assessment of the degree of interdisciplinarity reached, therefore this methodology can be iterative.

To sum up, in this paper we highlighted the importance of interdisciplinarity to effectively address the social implications of AI systems; we emphasized the need for a process that ensures that an interdisciplinary approach is properly adopted by research groups; we designed a methodology based on a narrative approach to ensure that a proper common understanding is reached, therefore making interdisciplinary design possible (Fig. 1).

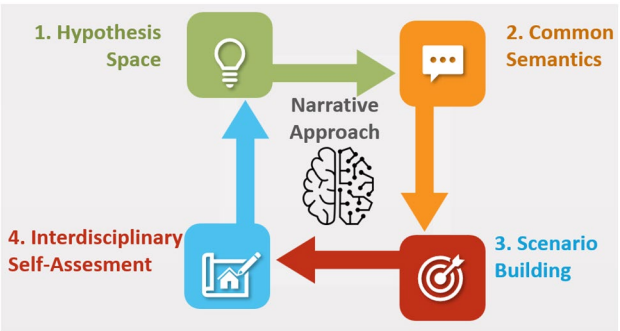

Fig. 1 The methodology process

Author contributions Concept: PB. Revision: PB, FM, DO. Proofreading: DO, HE, PB, FF, MP, FB, MF, LM, MC, FM, CS. Section I: FM, DO, MF, LM. Section II: MC, HE, CS. Section III: PB, LM, FF. Section IV: PB, MC, FM. Section V: PB, FB, FF, MP. Section VI: PB, DO.

Funding Open access funding provided by Scuola Superiore Sant'Anna within the CRUI-CARE Agreement.

References

High-Level Independent Group on Artificial Intelligence (AI HLEG) (2019) Ethics Guidelines for Trustworthy AI. Brussels

Angwin J, Jeff L, Surya M, Lauren K (2016) Machine Bias There's Software Used across the Country to Predict Future Criminals. And It's Biased against Blacks. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing (last accessed 28/05/2022)

Beaudouin V, Isabelle B, David B, Stéphan C, Florence d'A-B, James E, Winston M, Pavlo M, Jayneel P (2020) Flexible and context-specific AI explainability: a multidisciplinary approach. Preprint available at SSRN 3559477.

Blockeel H (2011) Hypothesis space. Encycloped Mach Learn 1:511–513

Christen M, Mark A, Salardi S, Saporit M (2020) A framework for understanding and evaluating moral technologies. In: Salardi S, Saporit M (eds) Le tecnologie 'morali' emergenti e le sfide etico-giuridiche delle nuove soggettività. Giappichelli Editore

Christin A, Alex R, Danah B (2015) Courts and predictive algorithms. *Data & CivilRight*. Available at https://www.law.nyu.edu/sites/default/files/upload_documents/Angele%20Christin.pdf (Last accessed 28/05/2022)

Chumkamon S, Hayashi E, Koike M (2016) Intelligent emotion and behavior based on topological consciousness and adaptive resonance theory in a companion robot. Biol Inspired Cognit Architect 18:51–67. https://doi.org/10.1016/j.bica.2016.09.004

Coeckelbergh M (2018) How to describe and evaluate 'deception' phenomena: recasting the metaphysics, ethics, and politics of ICTs in terms of magic and performance and taking a relational and narrative turn. Ethics Inf Technol 20(2):71–85. https://doi.org/10.1007/s10676-017-9441-5

Coeckelbergh M, Reijers W (2016) Narrative technologies: a philosophical investigation of the narrative capacities of technologies by using Ricoeur's narrative theory. Hum Stud 39(3):325–346

EU Commission (2021) Proposal for a regulation on a European approach for artificial intelligence. Brussels

Dignum V, Matteo B, Cristina B, Maurizio C, Raja C, Louise D, Gonzalo G, Galit H, Malte SK, Maite L-S, Roberto M, Juan P, Marija S, Matthijs S, Marlies van S, Stefano T, Leon van der T, Serena V, Tristan de W (2018) Ethics by design: necessity or

curse? In: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society (AIES '18). Association for Computing Machinery https://doi.org/10.1145/3278721.3278745
Dwivedi YK, Hughes L, Ismagilova E, Aarts G, Coombs C, Crick T, Williams MD (2021) Artificial Intelligence (AI): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy. Int J Inf Manag
Ebers M, Gamito MC (2020) Algorithmic governance and governance of algorithms: legal and ethical challenges. Springer, Cham. https://doi.org/10.1007/978-3-030-50559-2
Fisher WR (2021) Human Communication as Narration: Toward a Philosophy of Reason, Value, and Action. Univ of South Carolina Press, Durham
Flores AW, Kristin B, Christopher TL (2016) False positives, false negatives, and false analyses: a rejoinder to machine bias: there's software used across the country to predict future criminals. And it's biased against blacks. Fed. Probation 80(2). https://www.uscourts.gov/federal-probation-journal/2016/09/false-positives-false-negatives-and-false-analyses-rejoinder
Friedman B, Hendry D (2019) Value sensitive design: shaping technology with moral imagination. Mit Press, Cambridge
Friedman PB, Kahn H, Borning A (2020) Value sensitive design and information systems. Routledge, In The Ethics of Information Technologies
Green B, Chen Y (2021) Algorithmic risk assessments can alter human decision-making processes in high-stakes government contexts. In: Proceedings of the ACM on human–computer interaction
Henshel RL (1982) The boundary of the self-fulfilling prophecy and the dilemma of social prediction. Br J Sociol 33(4):511–528
High-Level Expert Group on Artificial Intelligence (2020) Assessment List for Trustworthy Artificial Intelligence (ALTAI) for Self-Assessment. Brussels
Kehl DL, Samuel AK (2017) Algorithms in the criminal justice system: assessing the use of risk assessments in sentencing. Responsive Communities Initiative, Berkman Klein Center for Internet & Society, Harvard Law School
Manders-Huits N (2011) What values in design? The challenge of incorporating moral values into design. Sci Eng Ethics 17(2):271–287
Mitcham C (1995) Ethics into Design. Explorations in Design Studies The University of Chicago Press, Discovering Design
Mumm J, Mutlu B (2011) Human-robot proxemics: Physical and psychological distancing in human-robot interaction. HRI 2011 - Proceedings of the 6th ACM/IEEE International Conference on Human-Robot Interaction 331–338. https://doi.org/10.1145/1957656.1957786
Mutlu B, Yamaoka F, Kanda T, Ishiguro H, Hagita N (2009) Nonverbal leakage in robots: communication of intentions through seemingly unintentional behavior. In: Proceedings of the 4th ACM/IEEE International Conference on Human Robot Interaction - HRI '09. https://doi.org/10.1145/1514095.1514110
European Parliament resolution of 20 October 2020 on intellectual property rights for the development of artificial intelligence technologies (2020/2015(INI))
European Parliament resolution of 20 October 2020 with recommendations to the Commission on a civil liability regime for artificial intelligence (2020/2014(INL))
European Parliament resolution of 20 October 2020 with recommendations to the Commission on a framework of ethical aspects of artificial intelligence, robotics and related technologies (2020/2012(INL))
Preece A, Dan H, Dave B, Richard T, Supriyo C (2018) Stakeholders in explainable AI. arXiv:1810.00184
Quattrocolo S (2020) Artificial intelligence, computational modelling and criminal proceedings. Springer, Cham
Quick KS (2018) The narrative production of stakeholder engagement processes. J Plan Educ Res. https://doi.org/10.1177/0739456X18791716
Reijers W, Mark C (2020) A narrative theory of technology. pp 79–111 in *Narrative and Technology Ethics*. Springer Cham.
Reijers W, Gordijn B (2019) Moving from value sensitive design to virtuous practice design. J Inf Commun Ethics Soc 17(2):196–209. https://doi.org/10.1108/JICES-10-2018-0080
Rudin C (2019) Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell 1(5):206–215
Seibt J, FlensborgDamholdt M, Vestergaard C (2018) Five principles of integrative social robotics. Front Artif Intell Appl 311:28–42. https://doi.org/10.3233/978-1-61499-931-7-28
Sharkey A, Sharkey N (2012) Granny and the robots: ethical issues in robot care for the elderly. Ethics Inf Technol 14(1):27–40. https://doi.org/10.1007/s10676-010-9234-6
Steinhaeusser SC, Schaper P, Bediako Akuffo O, Friedrich P, Ön J, Lugrin B (2021) Anthropomorphize me! Effects of Robot Gender on Listeners' Perception of the Social Robot NAO in a Storytelling Use Case. In Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction
van de Poel I (2020) Embedding values in artificial intelligence (AI) systems. Mind Mach 30(3):385–409
Van den Hoven J, Pieter EV, Van de Ibo P (2015) Handbook of ethics, values, and technological design: sources, theory, values and application domains. Springer, Cham
Wachter S, Brent M, Chris R (2017) Counterfactual explanations without opening the black box: automated decisions and the GDPR. Harv. JL & Tech. 31(2)
Wynsberghe V, Aimee. (2013) Designing robots for care: care centered value-sensitive design. Sci Eng Ethics 19(2):407–433