Design Research Society
DRS Digital Library

DRS Biennial Conference Series DRS2022: Bilbao

Jun 25th, 9:00 AM

Towards a living lab for responsible applied AI

Maaike Harbers
Rotterdam University of Applied Sciences, Netherlands, The

Anja Overdiek
Rotterdam University of Applied Sciences, Netherlands, The

Follow this and additional works at: https://dl.designresearchsociety.org/drs-conference-papers

 Part of the Art and Design Commons

Citation

Harbers, M., and Overdiek, A. (2022) Towards a living lab for responsible applied AI, in Lockton, D., Lenzi, S., Hekkert, P., Oak, A., Sádaba, J., Lloyd, P. (eds.), *DRS2022: Bilbao*, 25 June - 3 July, Bilbao, Spain. https://doi.org/10.21606/drs.2022.422

This Research Paper is brought to you for free and open access by the DRS Conference Proceedings at DRS Digital Library. It has been accepted for inclusion in DRS Biennial Conference Series by an authorized administrator of DRS Digital Library. For more information, please contact dl@designresearchsociety.org.

DRS 2022
Bilbao 25th June — 3rd July

Towards a living lab for Responsible Applied AI

Maaike Harbers*, Anja Overdiek

Rotterdam University of Applied Sciences, The Netherlands

*corresponding e-mail: m.harbers@hr.nl

doi.org/10.21606/drs.2022.422

Abstract: AI ethics research has mainly focused on high-level principles and guidelines, and technical issues. This position paper argues that more attention should go to the practical and contextual aspects of designing AI applications and explores how living labs can contribute to the ethical design, development and deployment of AI. Literature on AI ethics is discussed, and the term 'Responsible Applied AI' (RAAI) is introduced to refer to the ethical application of AI. Five requirements for the development of RAAI in a living lab are distinguished. Subsequently, the paper brings together literature from Open Innovation and Human Computer Interaction to examine the suitability of different types of living labs for developing RAAI. It concludes that Innovation Spaces (online and physical) combined with temporary and ethically governed Instrumented Places and People could be a fruitful environment for a living lab for RAAI. Implications and challenges for further research and practice are discussed.

Keywords: AI ethics; responsible AI; living lab; innovation space

1. Introduction

With the rise of artificial intelligence (AI) and its increasing impact on the world, attention for the ethics of AI has been growing. In the last years, at least 84 guidelines for AI ethics have been published (Jobin et al., 2019) and several articles appeared with overviews and reflections on work on AI ethics (e.g., Hagendorff, 2020; Jobin et al., 2019). One of the most important observations regarding this work is that most of it consists of abstract, high-level principles and guidelines, and that more research is needed to translate these into concrete tools, methods, and practices for designing, developing and deploying AI systems in specific application contexts (Ayling & Chapman, 2021; Mittelstadt, 2019; Morley et al., 2020; Hagendorff, 2019; Krijger, 2021). In this paper, we use the term 'Responsible Applied AI' (RAAI) as a response to this challenge, where 'responsible' refers to taking ethical considerations into account, and 'applied' marks the focus on the practical and contextual aspects of AI applications.

In the field of Human Computer Interaction (HCI), living labs have often been used to research, develop and experiment with new technologies (Følstad, 2008; Pallot & Pawar, 2012; Schuurman et al., 2019). As living labs are particularly suitable to research new applications

This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International Licence.

of technology when the fit of a technology to a specific context is significant (Haukipuro & Väinämö, 2019), they seem a promising approach for developing RAAI. Despite the large body of work on the use of living labs in general, little research had been done on the use of living labs as an approach for developing ethical or responsible AI specifically. The Netherlands AI Coalition (NLAIC) (https://nlaic.com) is advocating the use of ELSA labs (Ethical Legal Societal Aspects) to develop and deploy 'human centric AI', in line with the European focus on AI applications that respect fundamental rights and public values. A recent position paper published by the NLAIC provides a historical review of the notion of ELSA labs and an analysis of the critical debates around it (Van Veenstra et al., 2021). However, though the NLAIC position paper discusses concepts that are central to living lab approaches (e.g., co-creation and n-tuple helix), it does not explicitly refer to living labs.

In this position paper, we will explore how living labs can contribute to the design, development, and deployment of RAAI. First, in section 2, we will provide an overview of current work on AI ethics and argue that there is a need to shift the focus within this work to RAAI. Also, we will identify several requirements for an environment in which to practice RAAI. Second, in section 3, we will provide an integration of the literature on living labs from the Open Innovation and the HCI field. We will use a critical discussion of the five strands of living labs as distinguished by Alavi et al. (2019) to examine how different types of living labs cater to the RAAI requirements introduced in section 2, which leads to determining the best type of living lab for RAAI. Finally, in section 4, we will discuss the results of our analysis and their implications, and in section 5, we will end the paper with a conclusion.

2. The need for Responsible Applied AI

In the 1950s, AI started as a field of research that studied the development of computational systems able to perform tasks requiring human intelligence. For a long time, research in this field had a limited number of real-world applications. Currently, AI techniques are used in all kinds of applications, from social media algorithms to self-driving cars, and the range of AI applications is still growing. With the rise of AI, its impact on the world (societies, individuals, companies, nature, etc.) has increased, and with it, calls for research on the (ethical) implications and concerns around AI.

AI is applied for many different purposes and in a wide variety of domains and, therefore, knows multiple and diverse ethical issues. Some of the well-known concerns are: AI systems that threaten physical safety, like self-driving cars (Favarò et al., 2017) or autonomous weapon systems (Crootof, 2015); the employment of algorithms that unjustly advantage certain groups of people over others, e.g., based on their race, gender, or religion (Ntoutsi et al., 2020; West et al., 2019; Zuiderveen Borgesius, 2018); and AI's role in the spread of fake news and one-sided information (Agarwel et al., 2019; Cinelli et al., 2021). These issues are by no means an exhaustive overview of AI concerns, but they show that AI's impact on the world is considerable and that not properly addressing (ethical) issues around AI can result in significant harms.

2.1 Ethics and AI

The growing interest in the ethics of AI is visible in a number of ways. New research institutes around the topic have been founded, for instance, the AI Now Institute at New York University (https://ainowinstitute.org) and the Human-Centered Artificial Intelligence Institute at Stanford University (https://hai.stanford.edu). Also, new research communities have formed, e.g., the Fairness, Accountability, and Transparency in Machine Learning (FAT ML) community (https://www.fatml.org). Perhaps most visible, a huge number of ethical principles and guidelines on ethical AI have been proposed (Jobin et al., 2019). These principles and guidelines have been introduced by companies (e.g., Microsoft, Deepmind, IBM), governments (e.g., EU guidelines for trustworthy AI, Beijing AI principles, US Report on the Future of AI) and other organizations (e.g., OECD principles on AI, IEEE). There are several articles that provide comparisons and overviews (Zeng et al., 2018; Fjeld et al., 2020; Jobin et al., 2019; Hagendorff, 2020), of which Jobin and colleagues (2019) give the most extensive overview, comparing 84 principles and guidelines.

Although the above initiatives help to focus the discussion on ethics of AI, they are not sufficient to ensure ethical application of AI (Krijger, 2021). To make this argument, work on AI ethics can be organized along two dimensions: theory versus practice and technical versus non-technical focus.

Theory versus practice

Most of the existing principles and guidelines are defined on an abstract level, and multiple scholars point out that these do not provide sufficient guidance to design, develop, and deploy AI systems in practice (Ayling & Chapman, 2021; Mittelstadt, 2019; Morley et al., 2020; Hagendorff, 2019; Krijger, 2021). Little research has been done to study the effect of these theories on practice (Krijger, 2021; Ayling & Chapman, 2021). One of the few empirical studies performed on this topic found that reading ACM's code of ethics had virtually no effect on developers' ethical decisions (McNamara, 2020). Several authors argue that additional work is needed to translate the high-level guidelines and principles into concrete solutions, methods, tools and practices (Ayling & Chapman, 2021; Mittelstadt, 2019; Morley et al., 2020; Hagendorff, 2019; Krijger, 2021). In the words of Morley et al. (2019), a move is needed "from what to how".

Technical versus non-technical focus

The second dimension for organizing work on AI ethics involves the extent to which contributions in AI ethics focus on technical solutions versus non-technical solutions. In any real-world application of AI, the AI system is situated in a specific context and in a particular environment. Many, if not all, ethical issues around AI arise due to the interaction of these systems with their environment. Therefore, studying ethical issues of an AI system requires that the system is considered in its context of use (or its interactional context, see Dourish 2004)

and in society at large, or in other words, the sociotechnical system of which the AI system is part should be considered (Crawford & Joler, 2018; Dolata et al., 2021). In practice, however, not all work on AI ethics embraces this sociotechnical perspective. A recent literature study shows, for example, that the majority of articles tackling the issue of algorithmic fairness focus on technical solutions, such as, finding the algorithm that yields equal rates of prediction errors for different groups of people (Dolata et al., 2021). Fairness, however, is a social construct and lived experience, and different situations may require different notions of fairness and the perception of what is fair may change over time. A non-technical solution for achieving algorithmic fairness could, for example, involve developing a mechanism in an organization for ensuring the ongoing monitoring of algorithms regarding fairness.

Table 1. Categorization of AI ethics work along the dimensions of 1) high-level principles vs concrete practices, and 2) technical focus versus non-technical focus

	Technical focus		**Non-technical focus**
High-level principles	Ethical principles about issues for which technical fixes (seem to) exist (e.g., transparency)	«	Ethical principles about issues that for which pure technical fixes do not exists (e.g., ecological costs)
Concrete practices	Toolkits focusing on technical aspects (e.g., fairness of the data in a dataset)	«	Toolkits focusing on non-technical aspects (e.g., supporting external stakeholders in understanding the impact of AI)

Table 1 shows an overview combining the two dimensions of 1) high-level principles vs concrete practices, and 2) technical focus versus non-technical focus, described above. First, looking at the row or high-level AI ethical principles, principles addressing issues for which technical fixes seem exist get more attention than principles for which mere technical fixes do not exist. Hagendorff (2020) compared 22 AI ethics guidelines and provided an overview of the different issues they cover. He noticed that issues that recur in most of the guidelines (e.g., accountability, privacy, or fairness) are more easily solved by technological fixes than issues that are absent in most of the guidelines (e.g., political abuse of AI systems, lack of diversity in the AI community, and 'hidden' social and ecological costs). The issues in the second category are no less impactful, so it could be said that they are underrepresented in most of the current work on AI ethics.

Second, looking at the row of concrete practices, there seem to be more ethical tools and methods that focus on the technical aspects of the development of AI systems than on social, organizational, and legal aspects surrounding these systems. Wong and colleagues (2022) recently conducted a qualitative analysis of 27 AI ethics toolkits in which they found that most toolkits frame the work of AI ethics to be technical work to be performed by technical professionals. Social, organizational, and political aspects of AI ethics are not addressed in most of these toolkits (Wong et al., 2022).

As argued above, previous work has mostly focused on the upper left corner of the table: high level principles that seem to be addressable by focusing on technical aspects of AI. The least attention has gone to work in the lower right corner: concrete practices and methods for supporting non-technical aspects of AI ethics. In this paper, we argue that all four parts of the table are important. This means that developing methods for practicing AI ethics deserves much more attention than it currently gets. In addition, collaborating in order to develop these methods should consider the sociotechnical system of which an AI system is part, rather than only focusing on the technical aspects of AI. To fill in this gap, we argue for the design, development, and deployment of RAAI.

2.2 Requirements of Responsible Applied AI

As stated in the introduction of this paper, we will explore how living labs as an environment for social collaboration can contribute to the design, development, and deployment of RAAI. Before we turn to living labs in the next section, we identify five requirements for researching and practicing RAAI that a living lab should satisfy.

First, because of AI's large impact on society, RAAI should not only address effects of AI on individuals but also on society as a whole. Therefore, in line with the Dutch ELSA lab approach mentioned in the introduction (Van Veenstra et al., 2021) and the European Union's approach to AI (European Commission, 2020), we believe that the living lab should allow for addressing personal and public values. This means that design of RAAI should not only account for effects on users and other stakeholder groups (personal values), but also for effects of AI systems on society and the public good (public values), including long-term effects of AI, effects on sustainable development goals, and effects of possible abuse of AI systems.

Second, to ensure that public values as well as personal values are addressed and the perspectives of a wide variety of stakeholders included, the involvement of multiple stakeholders is needed. Living labs often involve a triple helix (research, industry, and government), quadruple helix (adding citizens) or quintuple helix (adding nature). Depending on the specific challenge at stake, multiple perspectives within one of the n-tuple groups could be needed, e.g., to represent minority groups.

Third, merely including multiple stakeholders is not enough, as some stakeholders hold more power than others, e.g., because one stakeholder is not as digitally literate as another. This is particularly relevant for the design of AI systems, as minorities are often affected by AI in different ways than majorities. Also, AI technology is complex and not everybody has the same level of understanding of the technology. Therefore, to ensure that voices of all stakeholders are heard in the design process, horizontal co-creation is needed, for instance, by supporting stakeholders with convivial tools (Sanders & Stappers, 2012).

Fourth, real-life experimentation with prototypes, based on the outcomes of co-creation sessions, is needed. It is impossible to foresee all possible consequences and implications of a new technology application during co-creation, particularly with such a complex technology as AI where implications are the result of the interplay between technology and humans

(Dourish, 2004). Real-life experiments can help to elicit the so-called unknown unknowns (Jensen et al., 2017), including long(er)-term effects and structural changes on the sociotechnical system.

Fifth, it is important to realize that developing RAAI is never 'done', as AI systems change (due to algorithmic learning), technological possibilities change, organizations change, and values can shift over time. An application that at one point is considered responsible AI, may at a later point in time fail to do so. Therefore, a living lab for RAAI should also allow and cater for ongoing reflection (Krijger, 2021).

3. Towards a living lab for Responsible Applied AI

Living lab as an approach, methodology and environment has been practiced since the early 2000s (Bergvall-Kareborn & Stahlbrost, 2009). It has been researched and developed with both an Open Innovation (Leminen et al., 2012; Almirall et al., 2012) and a user innovation (Dell'Era, & Landoni, 2014; Brankaert & den Ouden, 2017) lens. As multi-stakeholder collaborations in projects of a more institutionalized character, living labs have been used and studied in multiple contexts such as sustainability, healthcare and smart city making (see e.g., the 150+ active living labs associated with European Network of Living Labs, www.enoll.org). There is consensus in the field that co-creation and experimentation in a real-life use context are defining characteristics of this form of collaboration aimed at innovation and learning (Steen & van Buren, 2017; Hossain et al., 2019). Co-creation ideally means horizontal cooperation (all stakeholders have decision power) between at least stakeholders from the triple helix (government, research, industry) and often from the quadruple (including civic society and citizens) or quintuple helix (including the natural environment, see e.g., Carayannis et al., 2012).

In the absence of research on dedicated living lab environments for RAAI (see section 1), we rely on the more developed literature on living labs in the field of Human-Computer Interaction (HCI). It is promising that there is a growing number of papers that call to re-imagine the intelligent systems design process by fostering relationships between AI developers and UX designers with the goal of HCI+AI collaborations to make intelligent systems more fair, accountable, and transparent (Abdul et al., 2018; Yang et al., 2018). In HCI, the notion and practice of living labs developed differently to the field of Open Innovation. At least in practical application, the real-life user context characteristic of living labs seems to prevail over the horizontal co-creation principle, a phenomenon that becomes apparent in a recent study by Alavi et al. (2019). In their analysis across 152 living lab related papers, they extracted five divergent strands of living lab environments with overlapping but distinct conceptual frameworks (figure 1). They labeled them as "Visited Places", "Instrumented Places", "Instrumented People", "Lived-in Places", and "Innovation Spaces", and all but one of these types of living labs foreground the testing of technology with users. Section 3.1 looks deeper into these types of environments, complements them with additional literature and discusses them against the RAAI requirements defined in the previous section.

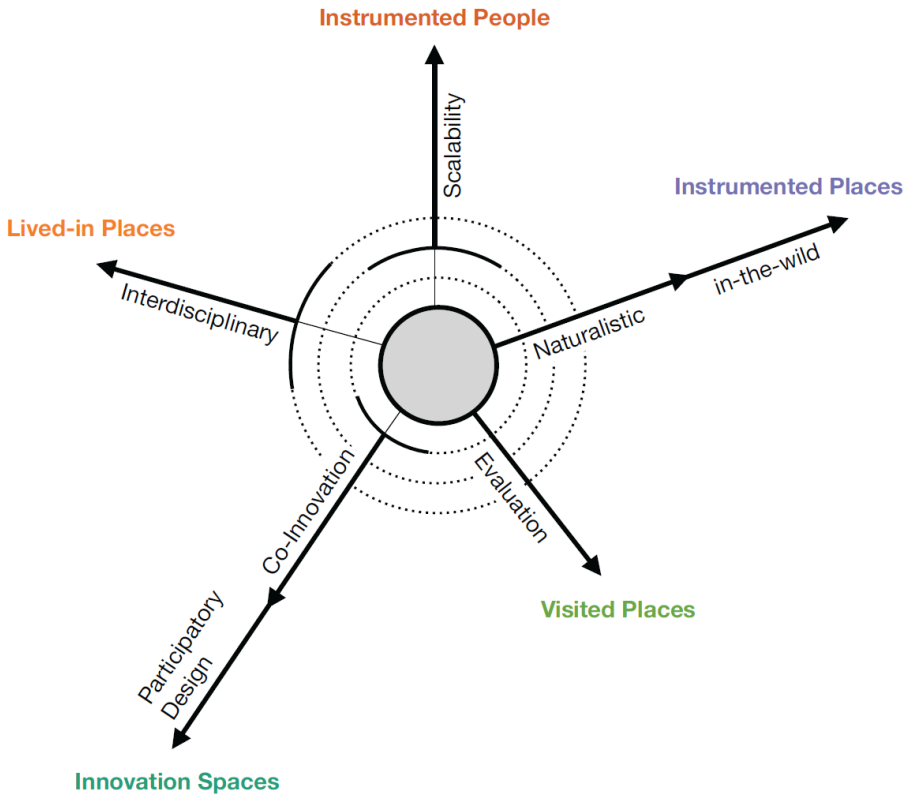

Figure 1 Living Lab Trends: the evolution of the core and new ideas leading to five trends of living lab. (Alavi et al., 2019, p. 12)

3.1 Five types of living labs as environments for Responsible Applied AI

Visited Places as environments rebuild typical living spaces (shops, living rooms, offices, etc.) and equip them with sensors. Visited Places are essentially laboratories for research, highly monitored and shaped to provide 'naturalistic' test environments. People are recruited to spend a limited time in these spaces while being exposed to experimental conditions. Visited Places resemble what Jones (2018) describes as a university lab geared to the conception, testing and evaluation of the possibilities of new technologies. They are suitable for evaluation purposes but lack multi-stakeholder co-creation as they are environments controlled by research. Gathering temporary experiences of people outside their everyday space, they also miss the real-life characteristics of a functional social context. The reflection of private or public values is rarely facilitated in Visited Places, let alone on-going reflection.

Like Visited Places, *Lived-in Places* are built for the specific purpose of facilitating research projects. However, unlike Visited Places, they are functional, built environments which are used as real apartments, offices, and so forth. The occupants of these environments are the permanent participants of studies. They are often engaged by low rents or even payment for their stay in the lab (see e.g. Taylor, 2020). Research in Lived-in Places can be confined to collecting data from the daily activities of people or entail the occupants' active commitment to giving feedback about a new technology or experience. Usually, different projects run

simultaneously in a Lived-in Place, with different professional and research disciplines working together, allowing for a high degree of interdisciplinarity. Laws and regulations are often wavered for the time of experimentation. Lived-in Places facilitate more than temporary testing with people but are still different from socially embedded real-life environments. This makes them less apt as a RAAI living lab. Although they are mostly triple helix collaborations, they rarely allow for horizontal co-creation as the contributing people typically do not decide about the projects' topics and goals. Lived-in Places also do not facilitate public value or long-term reflection on the social practices of technologies.

Instrumented Places on the other hand are real living environments with their ordinary businesses or inhabitants who agreed to participate in a study at their offices or homes and allow researchers to collect data for a certain period. By recruiting a community of individuals who, e.g., agree to carry a wearable device or install recording applications on their smartphones, the next type of living lab called *Instrumented People* provides researchers with highly scalable sources of data that are not necessarily bound to a physical location. The Instrumented People model creates opportunities for engaging users in addressing certain issues that may be situated in dispersed locations. For our goal, the two lab types can be merged into one type, *Instrumented Places and People*, as their only difference is the place-bound versus people-bound instrumentation.

Though Instrumented Places and People can be an effective platform for mobile sensing and large-scale contextualized data collection, there is rising criticism to these two strands of living labs. Taylor (2020) shows with the analysis of a planned Instrumented Place in The Netherlands that a growing number of this kind of lab "experiments on people, using technology" (Taylor 2020, p. 2), where they should rather experiment on things, using (informed) people. Alavi et al. (2019, p. 22-23) also stress the importance of trust and empowerment in relation to participating users and citizens in living labs (but without elaborating on how to achieve these).

Keeping in mind that for RAAI the informed partnership status of stakeholders and users participating in the lab is of great importance, the living lab type *Instrumented Places and People* could be a fit for a RAAI lab in the development phase. To study AI systems in their real-life context, Instrumented Places and People could work as temporary living labs, deployed for iteration and testing goals together with, e.g., companies or municipalities and their employees or citizens from a certain neighborhood. However, for the purpose of engaging different stakeholder in a more on-going dialogue about the effects and the co-evolution of AI systems with humans, a living lab environment which is more centered on horizontal co-creation and long-term engagement is needed.

The last type of living lab found by Alavi et al. (2019), *Innovation Spaces,* has emerged as a response to closed innovation environments and limited interaction of businesses with potential new markets. They promote the concept of "democratizing innovation". The vision of participatory design complements this framework by foregrounding the bottom-up long-

term collaborations amongst diverse stakeholders and by introducing a focus on socio-material working relations. Innovation Spaces bring together companies, research organizations, individuals and civic sectors as stakeholders; they typically take the shape of workshop rooms (ideation or maker spaces). Thus, Innovation Spaces facilitate and foster n-tuple cooperation.

In the HCI literature, Innovation Spaces are presented in relation to the need for engaging users in the early stages of the design process of a technology, but in principle they could also be used for on-going reflection. However, this would challenge the physical living lab to engage stakeholders over time. An Innovation Space, in contrast to the other types of living labs, is first and foremost a social environment that can meet social needs, create social relations, and within that social atmosphere examine participation in innovative creation. Personal values can be addressed, as well as public values. In Innovation Spaces, researchers typically draw on the principles of Participatory Design which should allow for horizontal co-creation. Participatory Design is also seen essential in the design of AI systems (Neuhauser et al., 2013), and can help to create ideas for AI applications through diverse methods, if users have (or get) a basic understanding of what AI can do and cannot do (Bratteteig & Verne, 2018). “However, to examine and understand the direct or unanticipated impact of the AI system requires investigating the human-AI system interactions” (Auerhammer, 2020, p. 1323). This would need to take place in another environment as Innovation Spaces do not facilitate real-life experimentation.

The four types of living labs derived from the discussion of Alavi et al. (2019) can be complemented by the recent and still under-researched area of “collaborative digital innovation tools” (De Moor et al., 2010; West & Bogers, 2014; Leminen & Westerlund, 2017). Haukipuro and Vainamo (2019) describe in a longitudinal case study how a community-based online platform can add value to a physical living lab of the type Innovation Space. In this case the platform engages a diverse community of users with interesting content (connected to the physical Innovation Space) and strong moderation. It is mainly used for collecting ideas and testing around digital applications and services, in all phases of the innovation process. Because of its efficient and remote functioning, the online platform can reach more divers users than the physical lab alone. The platform also allows for public and private areas and for anonymity, which could all be interesting features for an ongoing deliberation covering the use-time of AI. Adding an "Online Innovation Space” to the physical one could be particularly interesting for a living lab environment catering to the needs of the ongoing deliberation about AI system in the use-time. More research is needed to show if an *Online Innovation Space* can also facilitate horizontal co-creation.

Table 2 shows the critical discussion of this section in an overview. In the table, an ‘x’ means that the RAAI requirement is fulfilled by the living lab type, and an ‘(x)’ means that the requirement is partly satisfied. The overview shows that none of the living lab types alone fulfills all five RAAI requirements, but that they can be met with a combination of lab types. More specifically, Innovation Spaces (physical and online) combined with temporary and

ethically governed Instrumented Places and People, could be a fruitful environment for a living lab for RAAI. The combination of environments is the mark of a typical third generation living lab (Leminen et al., 2017) which works like a platform.

Table 2. Comparing Responsible Applied AI requirements to types of living labs

RAAI requirements	**Visited Places**	**Lived-in Places**	**Instrumented Places and People**	**Innovation Spaces**	**Online Innovation Spaces**
Personal and public values				x	x
n-tuple stakeholder collaboration		x	(x)	x	x
Horizontal co-creation				x	(x)
Real-life experimentation		(x)	x		
Ongoing reflection				(x)	x

Summarizing, the unique advantages of this combined lab for RAAI can best be clarified taking the perspective of a RAAI project which would make its way through the lab. First the Innovation Space would lay the base of engaging stakeholders from an n-tuple helix. As a physical and social environment, it helps create social relations, and within this atmosphere to examine participation in innovative creation. Personal values can be addressed, and a broader deliberation about public values can be organized. Also, rules of informed partnership and horizontal co-creation can be established. When the multi-stakeholder project is on its way past the research and conception phase, temporary labs (at the site of the participating partners or 'in the wild') would allow for the necessary real-life experimentation, and contextual testing of AI systems and lastly, after the implementation of the innovation, an Online Innovation Space could be used for ongoing reflection. Because of its efficient and remote functioning, the online platform could also reach more diverse stakeholders and users than the physical lab alone, spreading innovation results and engaging new stakeholders for the physical lab. Moreover, the online platform would allow for public and private areas and for anonymity, which could all be interesting features for an ongoing deliberation covering the use-time of AI.

4. Discussion

The findings of the previous section have several repercussions and present design challenges and topics for further research that will be briefly touched in this section. Firstly, the favorable lab environment for RAAI asks for the development of a matching methodology. This methodology will have to address how the interaction of the combined parts of the living lab can work. An approach that looks from the perspective of the different 'learning journeys' of the engaged stakeholders (researchers, companies, governmental partners, citizens)

could be very constructive here, as it would help to align and make sense of the different activities of the lab. Moreover, the combined RAAI living lab will have to be integrated with other research and education spaces. Jones' (2018) systemic analysis of the relationship between different "experimentation and learning spaces" according to their openness could be helpful here. The lab methodology will also need diverse ways to visualize, explain and communicate the workings of AI systems to stakeholders in the lab. There is extensive literature on designing tools for the early stages of innovation (e.g., Sanders & Stappers, 2012). For the much less researched stage of on-going reflections, RAAI could draw on the growing literature on Meta-Design and End-User Development (EUD) (Giaccardi & Fischer, 2008; Fischer et al., 2017) where participation and emergence are conceptualized as design spaces and IT systems are seen as living entities that can be evolved by their users. EUD research experiments with new ways to empower professional and non-professional users of IT systems with the skills to collaborate which could be integrated in RAAI living lab practice.

Finally, there are some important concerns regarding feasibility. The organization and maintenance of a living lab environment as proposed is a very resource-intensive endeavor. Next to financial investment in space, lab coordinators, communicators, and online community curators, it asks for new skills and commitment over time. The driving party (Leminen et al., 2012) as context of this RAAI living lab needs to be able to align purely profit-driven interest in development and implementation of AI systems with an AI ethics-driven approach; engage stakeholders over a longer period of time; pay attention to inclusion and power (imbalances) in horizontal co-creation to prevent participation "tokenism"; and overcome differences in language and barriers between different stakeholders, e.g., by using probes or boundary objects that make AI tangible. Moreover, due to its complexity explaining the value of investing time and other resources into the lab to stakeholders is a challenge. Finally, the lack of knowledge in the field of "collaborative digital innovation tools" makes any development of this kind of lab experimental and research-intensive.

5. Conclusion

This paper argued for more attention in AI ethics to the practical and contextual aspects of AI, introduced the concept of Responsible Applied AI (RAAI) along with five requirements for RAAI, and integrated AI ethics literature with literature on living labs from Open Innovation and HCI. Doing this, it contributes to theory by ordering and integrating literature from different fields. It also helps design researchers and practitioners working on AI ethics for their orientation. To explore how living labs can contribute to the practice of RAAI we mapped five types of living lab environments to the five RAAI requirements. A living lab environment specially catered to RAAI was proposed, consisting of a combination of a physical and digital Innovative Space, with temporary and ethically governed Instrumented Places and People. This paper thereby also helps institutions to realize the specific requirements of designing, developing, and deploying RAAI. It also assists with conceiving and designing spaces for this and prepares them for the complexity of such an endeavor.

Acknowledgements: This work has been funded by the NWO-SIA SPRONG grant 'Responsible Applied AI', grant number SPR.ALG.01.024.

6. References

Abdul, A., Vermeulen, J., Wang, D., Lim, B. Y., & Kankanhalli, M. (2018). Trends and trajectories for explainable, accountable and intelligible systems: An hci research agenda. In *Proceedings of the 2018 CHI conference on human factors in computing systems*, paper 582, 1-18.

Agarwal, S., Farid, H., Gu, Y., He, M., Nagano, K., & Li, H. (2019). Protecting World Leaders Against Deep Fakes. In *CVPR workshops* (Vol. 1).

Alavi, H. S., Lalanne, D., & Rogers, Y. (2019). The five strands of living lab: a literature study of the evolution of living lab concepts in HCI. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 27(2), 1-26.

Almirall, E., Lee, M., and Wareham, J. (2012). Mapping living labs in the landscape of innovation methodologies. *Technology Innovation Management Review*, 2(9), 12-18.

Auernhammer, J. (2020) Human-centered AI: The role of Human-centered Design Research in the development of AI, in Boess, S., Cheung, M. and Cain, R. (eds.), Synergy - DRS International Conference 2020, 11-14 August, Held online. https://doi.org/10.21606/drs.2020.282

Ayling, J., & Chapman, A. (2021). Putting AI ethics to work: are the tools fit for purpose?. *AI and Ethics*, 1-25.

Bergvall-Kareborn, B., & Stahlbrost, A. (2009). Living Lab: an open and citizen-centric approach for innovation. *International Journal of Innovation and Regional Development*, 1(4), 356-370.

Brankaert, R., & den Ouden, E. (2017). The design-driven living lab: a new approach to exploring solutions to complex societal challenges. *Technology Innovation Management Review*, 7(1), 44-51.

Bratteteig, T., & Verne, G. (2018, August). Does AI make PD obsolete? exploring challenges from artificial intelligence to participatory design. In *Proceedings of the 15th Participatory Design Conference: Short Papers, Situated Actions, Workshops and Tutorial.* Volume 2, 1-5.

Carayannis, E. G., Barth, T. D., & Campbell, D. F. (2012). The Quintuple Helix innovation model: global warming as a challenge and driver for innovation. *Journal of innovation and entrepreneurship*, 1(1), 1-12.

Cinelli, M., Morales, G. D. F., Galeazzi, A., Quattrociocchi, W., & Starnini, M. (2021). The echo chamber effect on social media. *Proceedings of the National Academy of Sciences*, 118(9).

Crawford, K., & Joler, V. (2018). Anatomy of an AI System. Retrieved November, 18, 2021. https://anatomyof.ai/.

Crootof, R. (2015). War torts: Accountability for autonomous weapons. *University of Pennsylvania Law Review*, Vol. 164, 1347-1402.

De Moor, K., Ketyko, I., Joseph, W., Deryckere, T., De Marez, L., Martens, L., & Verleye, G. (2010). Proposed framework for evaluating quality of experience in a mobile, testbed-oriented living lab setting. *Mobile Networks and Applications*, 15(3), 378–391.

Dell'Era, C., & Landoni, P. (2014). Living Lab: A methodology between user-centred design and participatory design. C*reativity and Innovation Management*, 23(2), 137-154.

Dolata, M., Feuerriegel, S., & Schwabe, G. (2021). A sociotechnical view of algorithmic fairness. *Information Systems Journal*.

Dourish, P. (2004). What we talk about when we talk about context. *Personal and ubiquitous computing*, 8(1), 19-30.

European Commission. (2020). White Paper on Artificial Intelligence-A European approach to excellence and trust. Com (2020) 65 Final.

Favarò, F. M., Nader, N., Eurich, S. O., Tripp, M., & Varadaraju, N. (2017). Examining accident reports involving autonomous vehicles in California. *PLoS one*, 12(9), e0184952.

Fischer, G., Fogli, D., & Piccinno, A. (2017). Revisiting and broadening the meta-design framework for end-user development. In *New perspectives in end-user development* (pp. 61-97). Springer, Cham.

Fjeld J, Achten N, Hilligoss H, Nagy A, Srikumar M (2020) Principled artificial intelligence: mapping consensus in ethical and rights- based approaches to principles for AI. Berkman Klein Center for Internet & Society, Cambridge

Følstad, A. (2008). Living labs for innovation and development of information and communication technology: a literature review. T*he Electronic Journal for Virtual Organizations and Networks Volume 10*, "Special Issue on Living Labs", August, 99-131.

Giaccardi, E., & Fischer, E. (2008). Creativity and Evolution: a meta-design perspective. *Digital Creativity*, 19(1), 19-32.

Hagendorff, T. (2020). The ethics of AI ethics: an evaluation of guidelines. Minds and Machines, Vol 30, pp. 99-120. Springer.

Haukipuro, L., & Väinämö, S. (2019). Digital user involvement in a multi-context living lab environment. *Technology Innovation Management Review*, 9(10).

Hossain, M. Leminen S., & Westerlund, M. (2019). A systematic review of living lab literature. J*ournal of Cleaner Production*, no. 213, 976-988.

Jensen, M. B., Elverum, C. W., & Steinert, M. (2017). Eliciting unknown unknowns with prototypes: Introducing prototrials and prototrial-driven cultures. *Design Studies*, 49, 1-31.

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence,* 1(9), 389-399.

Jones, P. (2018). Contexts of co-creation: Designing with system stakeholders. In *Systemic Design* (pp. 3-52). Springer, Tokyo.

Krijger, J. (2021). Enter the metrics: critical theory and organizational operationalization of AI ethics. *AI & SOCIETY*, 1-11.

Leminen, S., Rajahonka, M., & Westerlund, M. (2017). Towards third-generation living lab networks in cities*. Technology Innovation Management Review,* 7(11): 21-35.

Leminen, S., Westerlund, M., & Nyström, A. (2012). Living Labs as open-innovation networks. *Technology Innovation Management*, 2(9), 6-11.

Leminen, S., & Westerlund, M. (2017). Categorization of Innovation Tools in Living Labs. Technology *Innovation Management Review*, (7)1, 15–25.

McNamara, A., Smith, J., & Murphy-Hill, E. (2018). Does ACM's code of ethics change ethical decision making in software development?. In *Proceedings of the 2018 26th ACM joint meeting on european software engineering conference and symposium on the foundations of software engineering* (pp. 729-733).

Mittelstadt, B. (2019). Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence*, 1(11), 501-507.

Morley, J., Floridi, L., Kinsey, L., & Elhalal, A. (2020). From what to how: an initial review of publicly available AI ethics tools, methods and research to translate principles into practices. *Science and engineering ethics, 26(4), 2141-2168.*

Neuhauser, L., & Kreps, G. L. (2011). Participatory design and artificial intelligence: Strategies to improve health communication for diverse audiences. In *2011 AAAI Spring Symposium Series.*

Ntoutsi, E., Fafalios, P., Gadiraju, U., Iosifidis, V., Nejdl, W., Vidal, M. E., & Staab, S. (2020). Bias in data-driven artificial intelligence systems—An introductory survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery,* 10(3), e1356.

Pallot, M., & Pawar, K. (2012). A holistic model of user experience for living lab experiential design. In *18th International ICE Conference on Engineering, Technology and Innovation*, 1-15: IEEE.

Sanders, E. B. N., & Stappers, P. J. (2012). *Convivial toolbox: Generative research for the front end of design*. Bis.

Schuurman, D., Herregodts, A. L., Georges, A., & Rits, O. (2019). Innovation Management in Living Lab Projects: The Innovatrix Framework. *Technology Innovation Management Review*, 9(3).

Steen, K., & Van Bueren, E. (2017). The defining characteristics of urban living labs. *Technology Innovation Management Review*, 7(7).

Taylor, L. (2021). Exploitation as innovation: research ethics and the governance of experimentation in the urban living lab. *Regional Studies*, *55*(12), 1902-1912.

Van Veenstra, A. F., Van Zoonen, L., & Helberger, N. (2021). ELSA Labs for Human Centric Innovation in AI. Position paper. Netherlands AI Coalition (NLAIC).

West, J., & Bogers, M. (2014). Leveraging External Sources of Innovation: A Review of Research on Open Innovation. *Journal of Product Innovation Management*, 31(4), 814–831.

West, S.M., Whittaker, M., & Crawford, K. (2019). Discriminating Systems: Gender, Race and Power in AI. AI Now Institute. Retrieved from https://ainowinstitute.org/ discriminatingsystems.html

Wong, R. Y., Madaio, M. A., & Merrill, N. (2022). Seeing Like a Toolkit: How Toolkits Envision the Work of AI Ethics. arXiv preprint arXiv:2202.08792.

Yang, Q., Banovic, N., & Zimmerman, J. (2018). Mapping machine learning advances from hci research to reveal starting places for design innovation. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems,* paper 130, 1-11.

Zeng, Y., Lu, E., & Huangfu, C. (2018). Linking artificial intelligence principles. arXiv preprint arXiv:1812.04814.

Zuiderveen Borgesius, F. (2018). Discrimination, artificial intelligence, and algorithmic decision-making. Report published by Strasbourg: Council of Europe, Directorate General of Democracy.

About the Authors:

Maaike Harbers is professor Artificial Intelligence & Society at Rotterdam University of Applied Sciences. She researches how designers of AI-applications can account for the ethical and societal implications of their designs.

Anja Overdiek is professor Cybersocial Design at Rotterdam University of Applied Sciences. Her field is digital social innovation in sustainability transitions and its design method. She uses theoretical, design-led and action research methodology based on a More-Than-Human-Centered approach.