Second African Digital Humanism Summer School, Accra, Ghana, 31 August – 05 September 2026

A Companion Reader

Digital Humanism Initiative
Technology for Humanity

August 2026

Programme	*Monday 31 August 2026*	*Tuesday 01 Sept 2026*	*Wednesday 02 Sept 2026*	*Thursday 03 Sept 2026*	*Friday 04 Sept 2026*
2nd African Digital Humanism Summer School, Accra, Ghana	*Introduction - Foundations*	*Perspectives on AI and Society*	*Policy – From Above and From Below*	*Application Fields of Digital Transformation*	*Challenges today*
(07:00 onwards: breakfast) Morning session I 08:00– 09:15 am (short break 09:15-09:20 am)	* *Welcome:* by Chair **Francis Saa-Dittoh**, ISOC President **Maud Adjeley Ashong Elliot, Central U, UPSA, CMU, UDS** * *Introduction round participants*	* Lecture 4: **Peter Knees**: *Artificial Intelligence, & Where It Will/Might/Could/Should Go*	* Lecture 8: **Saa Dittoh:** *Integrating Indigenous Knowledge into AI Development in Africa*	* Lecture 10: **Victor De Boer:** *AI Knowledge Graphs for Decolonizing Cultural Heritage*	* Lecture 14: **Bukelwa Ngoqo:** *Digital Ubuntu: Multi-Sectoral Collaboration for AI Adoption in Africa*
Morning session II 09:20 – 10:35 am (break 10:35 – 11:00 am)	* Lecture 1: **Hannes Werthner & George Metakides:** *Digital Humanism: What and Why?*	* Lecture 5: **Maud Adjeley Ashong Elliot:** *Digital Technologies and Policies in the African Public Sector*	* Lecture 9: **Ricardo Baeza-Yates:** *Responsible AI – Issues and Challenges*	* Lecture 11: **Prosper Simbarashe Maguchu:** *A Human Rights Approach to Fight Corruption: The Role of Ubuntu*	* Lecture 15: **Adams Bodomo:** *Digital Activism: Linguistic Human Rights for the Promotion of Democracy in Africa*
Morning session III 11:00 – 12:15 pm	* Lecture 2: **Francis Saa-Dittoh:** *From Radio to AI: African Community-Driven Development of Sustainable Information Systems*	* Lecture 6: **Julia Neidhardt:** *Recommender Systems and their Ethics*	* Slot reserved for project work	* Lecture 12: **Gaétan Hains:** *Low-tech AI – Low Cost, Low Energy, High Social Value Applications*	* Closing Session: *Evaluation, Future Plans*
Lunch 12:15 – 01:30 pm					
Afternoon session I 01:30 – 02:45 pm (break 02:45 – 02:55 pm)	* Lecture slot 3: Round Table discussion: ***Africa and Digital Humanism*** (moderated by **Anna Bon**) * ***Use Cases*** *for project work*	* Lecture 7: *Software Engineering after AI* **Carlo Ghezzi** + **Panel Discussion** chaired by **Hans Akkermans**) (supported by ACM SIGSOFT)	* Free afternoon	* Lecture 13: **Carine Mukamakuza:** *AI for African Health*	(4) **Stanley Mukasa & Jesse Thornburg**: Solution Refinement and Showcase Preparation
Afternoon session II: Project work 02:55 – 04:00 pm **DigHum Innovation Boot Camp**	(1) **Stanley Mukasa:** Intro to Boot Camp, Thematic Areas, Criteria and Showcase	(2) **Stanley Mukasa & Jesse Thornburg:** Problem Discovery and Stakeholder validation		(3) **Stanley Mukasa & Carine Mukamakuza:** Solution Validation and Implementation Pathways	End of Summer School with **social dinner** and drinks **Saturday 05 Sept 2026:** DigHum Innovation Boot Camp
Evening			**Reception at Austrian Embassy**		**Showcase Day** 08:00 am – 04:00 pm

Website: https://euridice.eu/event/second-african-digital-humanism-phd-summer-school/ Tentative Programme, version: 27 Aug 2026

Foreword

Dear Reader:

The Second African Digital Humanism Summer School will be held in Accra, Ghana, from Monday 31 August 2026 to Saturday 05 September 2026,. It takes place at the premises of Central University (Summer School week Mon 31 August – Fri 04 September) and UPSA (Innovation Bootcamp, Saturday 05 September 2026).

For this occasion we have compiled a Companion Reader of recent articles connected to the Digital Humanism themes of the Summer School. Each lecturer has selected one salient article written by them as author or co-author. The order of the articles follows the lecture program that you can find on the previous page. We hope that you enjoy this Companion Reader in the preparation of the Summer School, and also that you find it a valuable resource afterwards.

For more information on the Summer School, please see our website at:
https://euridice.eu/event/second-african-digital-humanism-phd-summer-school/

Programme & Organization Committee
Second African Digital Humanism Summer School, Accra, Ghana, 31 August – 05 September 2026

This event falls under the patronage of:

unesco

The authors are responsible for the choice and presentation of views contained on this website and for opinions expressed therein, which are not necessarily those of UNESCO and do not commit UNESCO.

The Second Africa Digital Summer School is supported and made possible through generous contributions from various organizations:

Gallery of Speakers and Articles

Second African Digital Humanism Summer School, Accra, Ghana, 31 August – 05 September 2026

Digital Humanism Initiative
Technology for Humanity

SPEAKER

Second African Digital Humanism Summer School, Accra, Ghana, 31 August – 05 September 2026

Hannes Werthner

Hannes Werthner is emeritus professor of e-Commerce at the Faculty of Informatics, TU Wien. Prior to joining TU Wien, he had several professorships at Austrian and international Universities. His research is in several fields such as Decision Support Systems, e-Commerce, e-Tourism, Recommender Systems, and lately in Network Analysis and Text Mining. Besides research and teaching he is active in starting new initiatives, such as the Vienna PhD School of Informatics and the i2c (Informatics Innovation Center). In the area of e-Tourism, the International Federation for IT and Tourism (IFITT) grants the "Hannes Werthner Tourism and Technology Lifetime Achievement Award" to outstanding academics and/or professionals in the field.

Lecture 1a: *Digital Humanism – What and Why?*

Reader article:

Hannes Werthner: Digital Transformation, Digital Humanism: What Needs to Be Done.
In H. Werthner, C. Ghezzi, J. Kramer, J. Nida-Rümelin, B. Nuseibeh, E. Prem, A. Stanger (Eds.): Introduction to Digital Humanism – A Textbook, pp. 115-132. Springer Nature, Cham, Switzerland, 2024. https://doi.org/10.1007/978-3-031-45304-5_8

Digital Transformation, Digital Humanism: What Needs to Be Done

Hannes Werthner

Abstract Information technology (IT) changes our society and world, from the individual level up to the ongoing geopolitical powerplay. From an ontological point of view, it influences how we perceive the world and how we think about it. This transformation only happened in the short time span of about 80 years, and it is continuing. We highlight some of the major features of this process. Besides its enormous achievements, this development has serious shortcomings. We will discuss some of them and describe our positive answer: Digital Humanism, an approach that describes, analyzes, and, most importantly, influences the complex interplay of technology and humankind, for a better society and life, fully respecting universal human rights. It is a proactive approach, focusing on the integration of technical and social innovation. Then we will present our Digital Humanism initiative, discuss its research and innovation roadmap, and finish with a general framework integrating the different dimensions of this initiative. [Some arguments have already been partly expressed in other publications such as Werthner (Electronic Markets, 32:145–151, 2022a) and Werthner et al. (IEEE Computer, 56(1):138–142, 2023).]

1 Introduction

Today, we experience, also astonished by its transformative power, the complex technical socioeconomic process called *Digital Transformation*. This happened in a short time span, from the invention of the first electronic computer about 80 years ago to today, where IT acts as the operating system of our society. This metamorphosis from a stand-alone computer to a worldwide mega-machine touches on every aspect of our lives, and as stated by Lee (2020), we experience the co-evolution of man–machine.

But this progress, as the progress of our society, is full of contradictions, sometimes even pointing backward. As historical processes do not always move

H. Werthner (✉)
TU Wien, Wien, Austria
e-mail: hannes.werthner@tuwien.ac.at

H. Werthner et al. (eds.), *Introduction to Digital Humanism*,
https://doi.org/10.1007/978-3-031-45304-5_8

forward in a positive direction, this technological development also has its downsides, with negative impact in the economic, societal, and political arenas. It is a complex process, not easy to understand and to manage. Many forces play a role, technical, societal, political, and economic. One should add that economic interests should not play the decisive role. The historic claim of the *invisible hand* of Adam Smith, where following their self-interest consumers and firms create an efficient allocation of resources for the entire society, did not work, looking at the many distortions of this process. We only mention the ecological crisis and the increasing economic inequalities, major problems of today besides the accelerating worldwide power play on a political and even military level. Also in these global crises, IT plays an important role. In addition, in the Web itself, we observe an ongoing monopolization with few worldwide acting tech giants, which, besides exploiting (all) loopholes to avoid paying taxes, exercise economic as well as political power.

But technology is neither God given nor should it follow a technical or economic determinism. An approach to give society a voice, to demand for a democratic participation on equal terms, is Digital Humanism. As a concept and as an initiative with growing support, it looks critically at this development and tries to influence it.

In this contribution, we start by looking shortly at informatics[1] and one of its major successes, the Web, its development, as well as its broad and massive impact. We will then discuss some of the shortcomings of this technology-induced changes and present Digital Humanism and its Vienna Manifesto as a multidisciplinary and democratic answer. It is a proactive approach, focusing on the integration of technical and social innovation. One of its major contributions in this respect is the Digital Humanism research and innovation roadmap, serving as a guideline for future work, both practical and theoretical.

2 Some Notes on Informatics

Some notes on the nature of informatics will help explain its transformative power, the pervasiveness, the broad impact, and the speed of this development (for an excellent introduction to computer science, see the chapter by Larus). The computer as *the* machine of today is a general-purpose automaton. It can, as the sole automaton, control itself by software and also be instantiated by software to a particular specific problem-solving machine. This general-purpose machine has the unique property of being able to independently change and control its behavior based on external inputs and internal states. It is able to act on its own without human intervention, except the initial programming and setup. It demonstrates independent behavior. Stated differently, and which can be seen by the power of some AI tools, it automates and simulates human thinking.[2]

[1] In understanding, it is the basic science in the field, and though not totally correct, we will not distinguish between informatics and computer science.

[2] Or let's say, aspects of it

and they are the most valued companies worldwide.[6] They increase market efficiency through the reduction of transaction costs (Williamson, 1985). Focusing on transactional services, the big platforms are industrial sector independent. Concrete products play an almost negligible role; they are virtualized, as are companies, entire markets, and, increasingly, our society.

The current competition race in the field of AI also shows the power of these companies as well as the high concentration in the field. Only these companies and related startups are participating. The cost of training a very large AI system like ChatGPT and the associated requirements for computing power and data sets are concentrated in their hands. Academic institutions are less and less able to keep up.

At the end, we have a situation where these platforms offer services that have become public social goods and individuals and companies have to participate as not to be excluded from public life. In addition, given the plethora of information, we again need informatics and its intelligent tools to navigate the information space (Baeza Yates & Fayyad, 2022).

We add an economic observation: in contrast to the very optimistic promises of IT, we see that since the 1980s, which happens also to be the time of invention of the personal computer, income inequality has risen in practically all major advanced economies, parallel to the period of digitalization. But not only the gap within the society has widened; also, following the rules of the networked platform economy with its *winners take it all* principle, there is a growing market gap between companies. Furthermore, the productivity growth has slowed down, where one would have expected substantial growth, as forecasted by several public relations companies in the IT field. This is called the productivity paradox[7] by Nobel Prize winner Robert Solow, as though investments in IT grew, productivity did not react accordingly. Overall, the productivity growth rates halved since the 1980s, and also the labor's share of income fell significantly, accelerating after 2000 (Acemoglu & Restrepo, 2019).

It is not easy to isolate the causes of this socioeconomic development. In our view, there is a shared responsibility of a neoliberal ideology with a massive decrease in regulations, paralleled by ongoing concentration. Technology facilitated this process. Siddarth et al. (2021) even state that, referring to a growing economic consensus, one of the most important causes is the automation and labor replacement focus of technological change.

[6] Despite the current economic situation, the four most valuable companies in the world, w.r.t. market capitalization, are such platform companies, as of the end of 2022.

[7] See also Brynjolfsson (1993).

4 "The System Is Failing"

IT systems are both useful and extremely successful. When looking at the COVID-19 pandemic as an example, without IT tools, the world would have almost stopped, no work, no school, and no personal and public communications. In research, data science methods were essential for the development of effective vaccines. IT kept and keeps the system running, and it serves for solving fundamental and vital problems (see the essential role of informatics to tackle the Sustainable Development Goals (SDG) of the UNO).[8] Another, more recent example, although heavily discussed, is ChatGPT as an intelligent writing assistant, freeing the humans from tedious writing and formulating tasks. The importance of IT is also demonstrated by its important role on economic, political, and even military level. At the same time, this development comes with downsides, as stated already in 2018 by Tim Berners–Lee with his "The system is failing."[9] The list of critical and mutually dependent issues is long, and it is not complete (Werthner et al., 2023):

- *Concentration and monopolies* in the Web, where multinational IT companies have power that national governments have serious problems to control. These companies offer services that governments do not provide with that quality; they decide, and not the states, on the implementations of essential services for citizens, e.g., access to the mobile Web or cloud services.
- The centralization of power in the Web raises the issues of both *personal and geopolitical sovereignty* (Werthner, 2022b, or contribution of Timmers). These big companies decide on the implementations of crucial services. As an example, see the case of the Corona app in European countries, where big mobile app store providers (Apple and Google) decided on the architecture and functionality of these apps and not these countries.[10]
- *AI and automated decision-making*—put simply, the representation and automation of human thought may result in autonomous decision-making systems, with substantial legal and ethical questions (Larus et al., 2018). What makes it worse, in cases where AI is based on black-box algorithms, is we do not understand the outcome, i.e., decisions proposed and taken. AI tools like ChatGPT, based on a combination of both unsupervised training and reinforcement learning, simulate human mental abilities and conversation. They do this to an extent that some already see them as an existential threat to humanity. In addition, these tools are highly concentrated in the hands of a few powerful companies.

[8] sdgs.un.org/goals

[9] The Guardian, 12.03.2018

[10] Digital Humanism Online Lecture "Corona Contact Tracing – the Role of Governments and Tech Giants"; dighum.org/program-overview

- Further *automation* will have a massive impact on employment and jobs, in both qualitative and quantitative terms (chapter of Samaan). How are and will these new jobs be designed? Will machines augment and support humans, or will they replace them? In addition, the IT industry reproduces the colonial division of labor as described by Casilli (2021),[11] where much of the low-skilled work is done away from the rich metropolises (see also chapter of Munn). Or look at the so-called Gig economy with the mostly false self-assessment of the independence and freedom of work but actually mostly self-exploitation.
- Increasing *surveillance*, where we can observe violations of privacy on a massive scale, both by private companies and by state instances, was well described by Zuboff (2019) (see also chapter by Lindorfer). This is a major threat to liberal democracy. But whom do we trust: the big IT companies or governments? Here, civil society and democratic institutions will play a key role.
- In our online media, we see developments such as the intentional *fabrication of fake news* and the creation of *opinion bubbles* (chapter of Krenn & Prem). Originally intended for democratic and open communication and information exchange, these systems are increasingly becoming toxic in the political discourse and, consequently, a threat for democracy.
- Autonomous AI-based machines move to warfare, resulting in autonomous weapons. Already, UN Secretary General António Guterres states that "Autonomous machines with the power to . . . . take lives without human involvement. . . . should be prohibited by international law."[12] It is important to mention that here, the civil and academic society already reacted with the campaign to *Stop Killer Robots*, initiated by Topy Walsh to halt the development of autonomous weapons.[13]
- These developments in the IT domain create a substantial *environmental* burden. Although there are already a number of positive examples to use these tools for climate actions (e.g., increasing operational efficiency, proper data gathering and simulation, etc.), there is also quite a negative emission impacts of IT and its applications (e.g., development and training of tools on large data sets).[14] Also here, it depends on us, in which direction it goes.

[11] Cassili, A. (2021): Digital Humanism Lecture – What is a 'Truly Ethical' Artificial Intelligence? An end-to-end approach to responsible and humane technological systems. dighum.org/program-overview

[12] António Guterres on Twitter, Mar 25, 2019; 6:28 PM

[13] www.stopkillerrobots.org

[14] See our Digital Humanism Online Lecture by David Rolnick on "Is AI good or bad for the climate? ...it's complicated" (dighum.org/program-overview).

5 Digital Humanism and the Vienna Manifesto

This "double-sided" role of IT, its indisputable enormous achievements and potential, and at the same time its obvious critical issues, which are interlinked and connected, were the motivation for the first Vienna Workshop on Digital Humanism, in April 2019 (Werthner, 2022a). The intellectual point of departure was our responsibility as scientists (Popper, 1969), calling upon us to shape technologies according to human values and needs, instead of allowing technologies to shape humans. The workshop was also inspired by the tradition of the Vienna Circle, a multidisciplinary effort of the early twentieth century to reflect on the revolutionary implications of science for our understanding of the world (Sigmund, 2017).

Over 100 participants from academia, public institutions, civil society and business took part in the 2-day workshop. The program addressed the history and impact of IT and informatics, as well as the dynamics and future of the sector. The discussions focused on technical, political, economic, social, ethical, and legal issues. A real benefit was the presence of such a diversity of disciplines, covering political science, legal science, sociology, history, anthropology, philosophy, management science, and informatics. At the center of the discussion was the relationship between informatics and society, or, as expressed during the workshop, the co-evolution of information technology and humankind. The discussion showed that informatics alone, although important, is not enough to provide comprehensive answers; a much more broad, interdisciplinary approach is called for. The participants were also convinced that it is possible to influence these developments; indeed, that it is our responsibility to do so.

The term *Digital Humanism* was intentionally chosen to refer to the concepts of humanism and Enlightenment, according to which, the human is responsible for his or her actions and beliefs and is at the focus (Nida-Rümelin & Weidenfeld, 2018; see chapter of Nida-Rümelin and Staudacher). We also underline the importance of rational and critical reasoning, which is a reference to the Vienna Circle and its logical empiricism. We have the freedom, the right, and the responsibility to make use of our own thought power, and we are the authors of our own lives. Personal autonomy and freedom to make decisions are the prerequisites for an open, democratic, and environmentally sustainable society. Technological progress is not God-given nor does it follow a determinism. We, as individuals and as a society, should and must make decisions taking democratic, humanistic, and environmental considerations into account. We define *Digital Humanism as an approach that describes, analyzes, and, most importantly, influences the complex interplay of technology and humankind, for a better society and life, fully respecting universal human rights*.

At the workshop, a *Vienna Manifesto for Digital Humanism*[15] was discussed and adopted by signatories from nearly 50 countries as a blueprint for shared principles. The Manifesto is also a call to act collectively to mobilize support that transcends

[15] dighum.org/dighum-manifesto/

national borders and continents in order to build a more human and sustainable future. More specifically, the principles of the Vienna Manifesto include:

Privacy, democracy, and inclusion

- Digital technologies should be designed and deployed in such a way that they promote democracy and inclusion.
- Privacy and freedom of speech are basic values which should be at the center of our activities.

Regulation and public oversight

- The regulatory authorities must intervene to break up technology monopolies.
- Decisions whose consequences could affect individual or collective human rights must still be made by humans.

Specific role of science and the academic sector

- Scientific approaches integrating various disciplines and eliminating discipline-specific silos are needed for mastering our challenges.
- Universities are the places where new knowledge is created and critical thinking is exercised. They should eliminate the boundaries among disciplines and foster their collaboration toward a holistic view of technological development.

Education and training

- New curricula are required, which combine humanities, social sciences, and technical and engineering sciences.
- Education in IT and training work on the ethical and societal impacts of IT must begin as early as possible in the education process.

As one can see, Digital Humanism not only attempts to eliminate the downsides of this IT induced changes but to encourage human-centered innovation, and its focus is on making the world a better one to live in, to contribute to a better and sustainable society.

6 The Digital Humanism Initiative

We touched an obvious up-to-date and hot topic; the response was enormous. Not only academics from the informatics fields reacted but also from other disciplines, as was the expressed interest by the civil society, funding agencies, and political decision-makers.[16] In addition, there are a number of international initiatives with similar objectives with which we started to network, e.g., HAI or Human-Centered

[16]There was one very interesting result: the manifesto was featured in the Greek newspaper *Kathimerini*, and this article was then selected as key text for the Greek-wide university entry exams.

Artificial Intelligence at Stanford,[17] All Tech is Human,[18] Dutch Digital Society,[19] or the Digital Enlightenment Forum.[20]

The growing public awareness can also be seen in several recent political international actions, e.g., in the US antitrust lawsuits against Facebook and Google and in Europe *Digital Service Act, Digital Market Act*, proposal for *AI regulation*, or the European *GDPR*, all focusing on a regulation of the online world. Other signs to show that things are changing are, for example, the *OECD's principles on AI*, *UNESCO's* activities, *UN's* view on the Internet as global public good, or the global *Partnership on AI.* And international standardization organizations moved, e.g., IEEE with its *IEEE 7000 Software Engineering Standard*.

A specific reaction and another example of the growing public awareness came from Austrian governmental bodies and institutions. The Vienna government announced the foundation of an Institute on Digital Humanism, and the Vienna Science and Technology Fund (WWTF), a partner since the early beginnings, operates a Digital Humanism research program. The Austrian government signed the Poysdorf Declaration on Digital Humanism, a joint statement of the foreign ministers of Austria, Czech Republic, and Slovakia.

Due to the pandemic, our activities moved online with since then nearly 50 regular public online lectures and four workshops. This lecture series was a real success, and it contributed to the international discussion. It has a growing number of participants, internationally renowned speakers, and a wide range of topics from AI and ethics, limits of AI, or COVID-19 apps and privacy to the issue of sovereignty in the digital world.[21] In addition, we published the volume *Perspectives on Digital Humanism with* 49 contributions (Werthner et al., 2022) with currently already over 350,000 downloads,[22] produced a *Digital Humanism Roadmap for Research, Innovation and Teaching* (Prem et al., 2022), and organized a successful summer school.

The most important result is, however, that we succeeded in creating an intellectual core, consisting of the authors of the manifesto, the members of our international program committee, organizational partners (such as the bidt, def or L3S),[23] the authors of our different publications, and the speakers as well as the participants of

[17] hai.stanford.edu

[18] alltechishuman.org

[19] www.thedigitalsociety.info

[20] digitalenlightenment.org

[21] This material (already over 90 h videos) is online-accessible at dighum.ec.tuwien.ac.at/lectures-program and www.youtube.com/channel/UC-oCPW9l7IuDvu_J30tqMVw.

[22] As of June 6, 2023

[23] Bavarian Research Institute for Digital Transformation (bidt.digital), Digital Enlightenment Forum (digitalenlightenment.org), research center L3S (l3s.de)

our online Digital Humanism Lecture series. This really lively group is constantly growing and active (see also our statement on ChatGPT from March 2023).[24] This group represents our initiative, although it has no formal organization.

7 Research and Innovation Roadmap

Digital Humanism follows a constructive approach, its focus is on technologies that empower people, facilitate access to knowledge, enable participation and inclusion in society, and support diversity. We need to develop and deploy trustworthy systems, enabling the participation of varying stakeholders, which should augment humans. At the same time, our approach must address the discussed downsides of the digital transformation, make a system's decision transparent, detect malicious behavior, and provide privacy.

These objectives and constraints were the starting point for our Digital Humanism roadmap workshop, March 2022 (Prem et al., 2022). The participants discussed and produced the roadmap of Table 1. On the horizontal axis, you find the mentioned critical issues and on the vertical the research questions and tasks, the answering of which will contribute to the solution of the respective problems. Crosses in the cells indicate which research questions we consider essential for which critical issues.

This roadmap may serve as a starting point, as a basis to further discuss and define research and innovation programs in Digital Humanism. As examples, we look at three research questions:

- *Explainability:* How do we explain the decision proposed by a computer, and why is which content or product proposed? This is a hard issue in data-driven models, which have no explicit logic implemented; they can be seen as black boxes.
- *Fairness* is an important issue (such as AI, how fair or biased are training data), automation (fairness is a crucial concept when looking at the division of labor and the interaction of a system and human), platforms (how would a fair participation of clients look like), online media (e.g., what is a fair distribution of interesting online content for which user group), or environment (here one has often to deal with economic trade-offs; what is in this context a fair representation of the different interests?)
- *Efficiency and resilience*, where the latter refers to fault-tolerant systems and the surviving in and the fast recovery from critical situations. Obviously, distributed architectures, such as the Internet, and algorithms play a key role. This issue plays a role in the issues of automation, platforms, sovereignty, and environment.

[24] ChatGPT—a catalyst for what kind of future? dighum.ec.tuwien.ac.at/statement-of-the-digital-humanism-initiative-on-chatgpt

Table 1 Digital Humanism research and innovation roadmap (Prem et al., 2022), revised by P. Knees and H. Werthner

Research topics	Critical issues						
	AI and human control	Labor and automation	Surveillance	Platform and monopolies	Online media and fake news	Digital sovereignty	Environment and sustainability
Explainability	x				x		
Transparency	x		x		x		
Privacy	x		x			x	
Personalization	x		x	x	x		
Fairness	x	x		x	x		x
Norms and ethics	x	x	x	x	x	x	x
Accountability of systems and providers	x	x	x	x	x	x	x
Machine/human cooperation and control	x	x	x	x	x		
Participatory approaches	x	x			x		
Security			x			x	
Regulatory approaches	x	x	x	x	x	x	x
DigHum business models			x	x	x	x	x
Content moderation	x		x		x		
Market mechanisms and power relations		x		x		x	
Resilience and efficiency (algorithms and architectures)	x	x		x		x	x
Open systems engineering (incl. interoperability, open data)			x	x		x	

However, the roadmap is a first step, and it is not complete. It may not include all relevant research topics, and more importantly, it does not contain the different disciplines needed in each of the research topics. Let us take for example fairness, e.g., in recommendation or search. How is fairness defined? Is it with respect to the provider of information or products? Is it with respect to readers or consumers? Which subgroups? Do we need to define fairness with respect to some general societal criteria? Obviously, disciplines such as sociology, political science, and economics are needed. Consequently, Digital Humanism requires an exchange across various disciplines, throughout the entire process, i.e., when doing analysis, when developing new technologies, and when adopting them in practice. This is a challenge; it is hard, for instance, to come up with a common language, where all those involved use the same terminology with the same semantics. In addition, the way in which the research landscape is organized in separate silos still hinders interdisciplinarity. And interdisciplinary researchers, especially young ones, often have serious problems obtaining funding and support since they touch different communities but are not specialized enough in their core discipline. Interdisciplinarity has the danger to know too little about too many things.

This breadth represents also a challenge for teaching, how to integrate different disciplines, without losing scientific depth. Informatics departments worldwide have started to include topics such as ethics in their curricula, either as stand-alone courses or embedded in specific technical subjects. However, a real broad interdisciplinary curriculum covering the different aspects and disciplines is still missing.[25] But there are also positive developments, e.g., in systems engineering with steps to integrate ethical guidelines in the software process. Some companies already offer specific tools, and associations such as IEEE provide guidelines for ethical design of systems (Spiekermann-Hoff, 2021) (see chapter of Neppel and chapter of Zuber et al.).

In general, Digital Humanism calls for a different technology path. Instead of focusing on pure automation and optimization, we need to look at participation and to augment human capabilities. As explained earlier, such directions already existed in the early days of computing, for example, the work of Doug Engelbart with his foundational work in the field of human–computer interaction or his Augmentation Research Center Lab in SRI International or Vannevar Bush with his Memex concept as a tool to augment, not replace, humans.[26] This tradition in human-centered technology is continued by researchers like Ben Shneiderman with his human-in-the-loop approach (Shneiderman, 2022, also chapter of Sharp). In this context, one has to mention the interesting work going on in Taiwan, with its g0v community. It promotes transparency of government information and is committed to developing information platforms and tools for citizens to participate in society.

[25] Although there are some promising steps in this direction, e.g., the European project Aurora

[26] There were also first concerns regarding this technological development: already in the 1960s, Joseph Weizenbaum with his famous natural language processing program ELIZA simulating a conversation with a psychotherapist noted “powerful delusional thinking” about a system’s intelligence (Weizenbaum, 1976).

Audrey Tang, a core member of this community, became Taiwan's first Digital Minister, and the country has rolled out experiments in digital democracy, decentralized governance, and collective intelligence.[27]

We envisage a future based on a different way forward for the role of technology in a society that is inclusive and focusing on the needs of the human, the society, and nature. A promising approach is proposed by Siddarth et al. (2021), called *digital plurality*. It is (1) *complementary*, i.e., complement and cooperate and not replace; (2) *participative*, i.e., cooperative and co-evolving and respecting human rights; and (3) based on *mutualism*, i.e., heterogeneous approaches can benefit from each other. Technologies need to take into consideration social, political, economic, and ecological objectives. In the end, it is not about technology alone.

As the Internet and the Web have become a social public good, Digital Humanism as an initiative to shape technology according to human values is not about research and academia only. Assuming the concept of human technology co-evolution, this does not evolve on its own. As it is currently governed by unequal societal and economic power relationships, we also need to talk about power and politics. Thus, Digital Humanism needs a multidimensional framework, on three levels:

- Different problem areas as described in Sect. 4, from platforms to AI, privacy, or work
- Different disciplines: informatics and technical/engineering disciplines, social science and humanities; from analysis to construction, with, however, the real challenge of interdisciplinarity
- Different activities: applied and basic research, development and experiments, innovation, education, communication, and, finally, political intervention

This is complicated and it is challenging (Neidhardt et al., 2022), but it needs to be done.

8 Conclusions

We started by discussing the development of the Web and its transformation from an open information infrastructure to a centralized one. This comes along with several socioeconomic and political shortcomings, which I described; each of them is discussed in other chapters of this book. Digital Humanism as an approach that describes, analyzes, and, most importantly, influences the complex interplay of technology and humankind is a constructive answer to these developments. We ended by presenting our research and innovation roadmap as a guide for future work.

Digital Humanism takes a cross-disciplinary and ethical point of view. It also touches the political level, as at the end, the IT-induced economic and societal change is a political question. As there is no higher being that is responsible nor does these developments follow a historical determinism, we, the people, should be

[27] g0v.tw/intl/en/

the driving force, via democratic participatory approaches. We should not obey but manage and guide the process, especially as IT will not stop, nor will the changes it induces.

We should not focus on better and faster; we need a long-term and sustainable perspective. This is the lesson of Digital Humanism: looking at the achievements, opportunities, and threats of technology, it should serve for the better of a society. Let us use IT to create a socially and ecologically sustainable society. One statement of the Vienna Manifesto is that we should not only analyze and discuss but also act, both in practical and scientific terms. We are at a crossroads.

Discussion Questions for Students and Their Teachers

1. How are the critical issues of Sect. 4 interrelated?
2. Discuss the principles of the Vienna Manifesto. Are there any missing?
3. Discuss centralized vs decentralized developments of the Web.
4. Discuss some rows, i.e., research questions, of the roadmap, and identify the necessary disciplines for these research issues.

Learning Resources for Students

1. *Haigh, T, and Ceruzzi, P.E. (2021) A new History of Modern Computing. The MIT Press. 2021*

   From microchips to cellphones to gigantic server farms, computers are among history's most revolutionary and rapidly evolving technologies. Yet their own history is littered with myth, misunderstanding, and misinformation. Written by distinguished experts, this book tells the story of where computers came from, how they changed the world, and why those changes mattered to diverse communities. This book is essential to historians, curators, and interdisciplinary scholars in informatics, information, and media studies.
2. *Aiello, M. (2018) The Web Was Done by Amateurs. Springer. 2018*

   Divided into four parts, it critically reflects on the Web's historical path. It starts with the prehistory of the Web, describes the original Web proposal as defined in 1989 by Tim Berners–Lee, and the most relevant technologies associated with it. Then it combines a historical reconstruction of the Web's evolution with a more critical analysis of its original definition and the necessary changes made to the initial design. Finally, it reflects on its technical and societal success. It was written with a technologically engaged and knowledge-thirsty readership in mind.
3. *Mitchell, M. (2019). Artificial Intelligence: A Guide for Thinking Humans. Picador. 2019*

   The book stresses that computers lack the general intelligence that we, humans, have. The author argues that achieving superintelligence would require that machines acquire commonsense reasoning abilities that are nowhere in sight. The book also contains a worthy historical overview.
4. *Crawford, K, (2021) Atlas of AI. Yale University Press. 2021*

   Kate Crawford reveals how AI is a technology of extraction: from the minerals drawn from the earth to the labor pulled from low-wage information workers to the data taken from every action and expression. This book reveals how this

planetary network is fueling a shift toward undemocratic governance and increased inequity. Rather than focusing on code and algorithms, the author offers us a material and political perspective on what it takes to make AI and how it centralizes power.

5. *Cohen, J. (2020). Between Truth and Power. MIT Press. 2020*

   Profound analysis and thinking about the two-way interplay between corporate and government in policy making, building on the ideas of governmentalism of Foucault and "law is code" of Lawrence Lessig.

6. *Kelly, K (2016). The Inevitable: Understanding the 12 Technological Forces That Will Shape Our Future. Penguin Books. 2016.*

   In this book, Kelly talks about how Internet scale never would have been possible top–down. He compares TV networks against Internet content creation, positing that by recruiting the users, the latter is swamping the former. He argues that AIs will be distinctly nonhuman intelligences and will turn into multiple intelligence species. While McLuhan noted that tools are extensions of ourselves, Kelly notes that the cloud is an extension of our souls. Citing the "adhocracy" of Wikipedia, he observes that we don't need much top–down design to get fantastic outcomes. We only need a little.

7. *Shneiderman, B. (2022) Human-Centered AI. Oxford University Press. 2022*

   The focus is on the opportunities of AI and how it presents and how to exploit them. The author also puts forward 15 recommendations about how to implement human-centered AI and how to bridge the gap between ethical considerations and practical realities to make successful, reliable systems.

8. *Werthner, H., Prem, E., Lee, E.A., Ghezzi, C. (2022): Perspectives on Digital Humanism, Springer. 2022.*

   This open-access book contains essays by selected thinkers from computer science, law, humanities, and social sciences, reflecting on Digital Humanism, what it is, and what it wants to achieve. It serves as further introduction to this emerging field, and it sets an agenda for research and action.

References

Acemoglu, D., & Restrepo, P. (2019). Automation and new tasks: How technology displaces and reinstates labor. *Journal of Economic Perspectives, 33*(2), 3–30.

Baeza Yates, R., & Fayyad, U. (2022). The attention economy and the impact of artificial intelligence. In H. Werthner, E. Prem, A. Lee, & C. Ghezzi (Eds.), *Perspectives on digital humanism*. Springer.

Brynjolfsson, E. (1993). *The productivity paradox of information technology*. CACM Dec 1993, 36/12.

Hardin, G. (1968). The tragedy of the commons. *Science, 162*, 1243–1248.

Larus, J. , Hankin, C., Carson, S. G., Christen, M., Crafa, S., Grau, O., Kirchner, C., Knowles, B., McGettrick, V., Tamburri, D. A., & Werthner, H. (2018). *When computers decide: European recommendations on machine-learned automated decision making*. Joint report Informatics Europe & EUACM. www.informatics-europe.org/publications

Lee, E. A. (2020). *The coevolution. The entwined futures of humans and machines*. MIT Press.

Neidhardt, J., Werthner, H., & Woltran, S. (2022). It is simple, it is complicated. In H. Werthner, E. Prem, A. Lee, & C. Ghezzi (Eds.), *Perspectives on digital humanism*. Springer.
Nida-Rümelin, J., & Weidenfeld, N. (2018). *Digitaler Humanismus*. Piper.
Nygaard, K. (1986, September). Program development as a social activity. In: Kugler, H. J. (Ed.), *Information Processing 86, Elsevier Science, IFIP. Proceedings from the IFIP 10th World Computer Congress, Dublin, Ireland*.
Ostrom, E. (1990). *Governing the commons: The evolution of institutions for collective action*. Cambridge University Press.
Peterson, T. L., Ferreira, R., & Vardi, M. Y. (2023). Abstracted power and responsibility in computer science ethics education. *IEEE Transactions on Technology and Society*.
Popper, K. (1969, March). Moral responsibility of the scientist. *Encounter*.
Prem, E., Hardman, L., Werthner, H., & Timmers, P. (Eds.) (2022). *Research, innovation and education roadmap for digital humanism*. The Digital Humanism Initiative, Vienna. owncloud.tuwien.ac.at/index.php/s/vmZSxsuruhk77Iy
Shneiderman, B. (2022). *Human-centered AI*. Oxford University Press.
Siddarth, D., Acemoglu, D., Allen, D., Crawford, K., Evans, J., & Jordan, M. (2021, December). *How AI fails us*. Technology & Democracy Discussion Paper. Harvard University.
Sigmund, K. (2017). *Exact thinking in demented times: The Vienna circle and the epic quest for the foundations of science*. Basic Books.
Spiekermann-Hoff, S. (2021). What to expect from IEEE 7000TM – The first standard for building ethical systems. *IEEE Technology and Society Magazine, 40*(3), 99–100.
Stanger, A. (2020). Consumers vs. citizens in democracy's public sphere. *CACM, 63*(7), 29–31.
Vardi, M. (2018, July). How the hippies destroyed the internet. CACM, 61/7.
Vincent, N., & Hecht, B. (2021). A deeper investigation of the importance of wikipedia links to search engine results. *Proceedings of the ACM on Human-Computer Interaction, 5*(CSCW1), 1–15.
Weizenbaum, J. (1976). *Computer power and human reason: From judgment to calculation* (1st ed.). W H Freeman.
Werthner, H., Prem, E., Lee, A., & Ghezzi, C. (Eds.). (2022). *Perspectives on digital humanism*. Springer.
Werthner, H. (2022a). From absolute nonsense to the world's operating system. *Electronic Markets, 32*, 145–151.
Werthner, H. (2022b). Geopolitics, digital sovereignty. . .what's in a word? In H. Werthner, E. Prem, A. Lee, & C. Ghezzi (Eds.), *Perspectives on digital humanism*. Springer.
Werthner, H., Stanger, A., Schiaffonati, V., Knees, P., Hardman, L., & Ghezzi, C. (2023). Digital humanism: The time is now. *IEEE Computer, 56*(1), 138–142.
Williamson, O. (1985). *The economic institutions of capitalism*. Macmillan.
Zuboff, S. (2019). *The age of surveillance capitalism. The fight for a human future at the new frontier of power*. Public Affairs.

SPEAKER

Second African Digital Humanism Summer School, Accra, Ghana, 31 August – 05 September 2026

George Metakides

George Metakides is Professor Emeritus from the University of Patras, Honorary President of the Digital Enlightenment Forum and Advisor to several international organizations. He is often a Visiting Professor in European Universities. He is involved in the analysis of the economic, political and social impact of digitization, related cybersecurity, data protection and regulatory issues and the promotion of international cooperation towards a digital ecosystem respecting shared human values. With a Ph.D. in Mathematical Logic earned from Cornell University in 1971, he pursued an academic career at MIT, Cornell and Rochester University before returning to Greece as Chair of Logic at the University of Patras. Since 1984 he has held senior positions with responsibility for Research & Development policy, funding and international co-operation in European institutions including the Directorship of the ESPRIT program He has contributed to the establishment of international institutions (including the launch of the World Wide Web consortium in 1993), has received a number of awards and honorary degrees and is a corresponding member of several National Academies.

Lecture 1b: *Digital Humanism – What and Why?*

Reader article:

George Metakides: The Time to Democratize our Digital Future Has Come. IWM Post No. 133, p. 16, Spring/Summer 2024. Institut für die Wissenschaften vom Menschen / Institute for Human Sciences (www.IWM.at), Vienna, Austria. https://www.iwm.at/publication/iwmpost-article/the-time-to-democratize-our-digital-future-has-come

The Time to Democratize our Digital Future Has Come

© gguy / shutterstock.com

IWMpost 133: False Prophets, False Promises

Author: George Metakides

The recent stunning AI developments call for fast and vigorous action, free from the illusions of the turn of the 21st century, so that the benevolent potential of these technologies can start being harnessed for the collective good.

After the collapse of the Soviet Union, Western-type liberal representative democracy was no longer considered under threat and it started being taken for granted. The theory was put forward that liberal democracy was a "natural state" to be nurtured, preserved, and spread by market power and globalization. This created a sense of euphoria characterized emblematically by Francis Fukuyama's "end of history."

While the related geopolitical developments have been addressed by historians and political scientists, it is often overlooked that, practically concurrently, the World Wide Web was born

and began to blossom, further strengthening the euphoria of the 1990s as it appeared to pave the way for a digital golden age of democracy—a cultural renaissance that would reinvent democracy as a digital Athenian agora where goods as well as ideas would be freely exchanged. This in turn, it was believed, with substantial preliminary evidence (for example, the early hopes of the Arab Spring), would empower more direct and informed citizen participation in open democratic societies.

Alas, this vision of milk, honey, and digital democratic bliss turned out to be an illusion. The undeniable positive attributes of the Web came with an increasing number of negative ones. As democracy started backsliding worldwide, scepticism about the impact of the Internet started growing, leading eventually Tim Berners-Lee in 2019 to call for "global action to save the Web from political manipulation, fake news, privacy manipulation and other malign forces that threaten to plunge the world into a digital dystopia." His dystopia has now been further exacerbated by the advent of generative AI and its facilitation of AI-powered disinformation, mob-driven social network behavior, democracy-threatening polarization, and threats to children's mental health.

In fact, democracy and digital technology have lived "parallel lives" since the 1990s. The euphoria immediately after the collapse of the Soviet Union caused a weakening of the defense and promotion of democracy in what Timothy Snyder called a "unilateral moral disarmament." A very similar and simultaneous sense of euphoria and unbridled techno-optimism prevented the anticipation of some of the negative impacts of digital technologies and the emergence of Big Tech, with its gigantic-scale monetization of personal data and the potential of its platforms to be used to disrupt and corrupt democratic processes. Market utopianism and techno-utopianism went hand in hand and reinforced each other.

How times change! At the beginning of the 21st century, "digital activists" were justifiably worried about potential government control of the digital public sphere but they failed to anticipate the threat of control by a very small number of very large companies. It is ironic that today they justifiably call for vigorous government intervention to rescue us from domination and control by Big Tech.

At the start of the Arab Spring, the role of social media was glorified, including with proposals like giving the Nobel Prize to Twitter. Fast forward to a recent U.S. Senate hearing during which Senator Lindsay Graham told Meta's Mark Zuckerberg that he had "blood on his

hands."

The U.S. Federal Trade Commission (FTC) is now suing Big Tech companies in an effort to curtail their monopolistic power and accuses them of "surreptitiously re-writing their privacy policies to allow themselves to use consumer data for their AI product development." In the EU a similar drive is underway through the Digital Market Act, the Digital Services Act, and the recently approved AI act. This sets the stage for a "tempered techno-optimism" approach that, perhaps ironically again, may be helped by the hype around generative AI developments in terms of expectations as well as of fears.

Civil society and political decision-makers are more ready than ever to support an approach of regulation and public investment to minimize the risks that digital technologies, and AI in particular, entail and at the same time to help harness the beneficent potential of these technologies for the collective good.

They also increasingly appreciate the threats that unprecedented concentration of economic and political power in the hands of a few super-large companies implies. Together, the "Magnificent Seven" (Apple, Amazon, Google, Meta, Microsoft, NVIDIA, and the group of companies owned by Elon Musk), currently have a valuation of about $13 trillion, rivalling the sum of the GDP of the four largest European economies (Germany, France, the United Kingdom, and Italy). Their economic power is coupled with the political power to manipulate, herd, and polarize to an extent that the public sphere that is essential for liberal democracy to function properly is severely corrupted.

That the concentration of economic and political power represents a danger to democracy was recognized in the United States in the 1890s as articulated by Senator John Sherman (of Sherman Antitrust Act fame), which led to vigorous antitrust legislation resulting in the breakup of Standard Oil and, much later, AT&T. This spirit was weakened in the 1970s and has not recovered since. The abovementioned efforts in the United States and the EU are an effort to revive this spirit after redesigning antitrust legal tools so that they are suited to the evolving digital ecosystem.

The sensitization and awareness that the recent AI developments have triggered can help us not to be collectively duped again into the same passive techno-optimism that allowed, for example, the unregulated and, in many ways, catastrophic development of social network platforms.

They can also accompany the regulatory framework that the EU has pioneered and must now be promoted for the broadest possible worldwide adoption, with adjustments, leveraging the "Brussels effect." This must come with generous public investments that allow all companies as well as national and local governments globally to develop digital tools designed to assist people rather than to replace them.

These investments are absolutely needed to make available to all the three main prerequisites for innovative, human-centered AI research and development: computing power, multilingual and locally curated data, and human resources with the required expertise. Failing to provide investment to enable such a "democratization of AI futures" will mean not only surrendering to current technological domination but also leaving the power to determine future research directions exclusively in the hands of Big Tech. This will in turn mean the abandonment of any pretence of a democratic public sphere as the Big Tech companies will continue, to quote Tim Berners-Lee's March 2024 open letter, "exploiting people's time and data with the creation of deep profiles that allow for targeted advertising and ultimately control over the information people are fed".

In the history of technological development, the owners of innovating companies (from the railroad, steel, and oil barons onward) always enjoyed a "grace period" during which they made their own rules and earned big profits as the fruits of their innovation. Then came a time when society realized that such a grace period lasting beyond a certain point leads to exploitation, the destruction of market competition, and eventually harms collective well-being. At that point, society decided to take action in the form of regulation accompanied by public investment.

The digital oligarchs have had their grace period for too long. The time to democratize our digital future is now.

George Metakides is professor emeritus of logic at the University of Patras, co-founder and honorary president of the Digital Enlightenment Forum, and advisor to several international organizations. He was a Digital Humanism Fellow at the IWM in 2023–2024.

MORE FROM: IWMPOST ARTICLE

Veering Right While Talking Left

Author: Vito Laterza

Italy's recent history shows why today's populism cannot be mapped neatly onto left and right. From the 2018 anti-immigrant consensus to pandemic-era diagonalism and Giorgia Meloni's electoral victory, exclusionary politics has learned to speak in…

The Ukrainian "Harvard Miracle"

Author: Sašo Ordanoski

The Ukrainian Research Institute at Harvard University is one of the most successful initiatives of the Ukrainian diaspora, having started as a grassroots movement. How did it come into being at a time when there was no independent Ukraine on the…

This Is Not a "Migration Crisis": Asylum Accommodation and the Real Divide

Author: Paolo Novak

What if the “migration crisis” is not about migration at all? Asylum accommodation renders visible the spectacle of borders while bringing crisis into communities. However, the real divide is not between citizen and foreigner but between those who…

Peter Thiel’s Empty Gospel

Author: Wojciech Engelking

Peter Thiel’s lecturing in Rome on the Antichrist should be seen neither as amusing nor as terrifying. Rather, it ought to be regarded as a window through which one can see what, after more than one year of Trump’s rule, remains of the tech bros’…

Der Hegemon im Niedergang

Author: Manuela Boatcă

Aus einer globalen und historischen Perspektive, wie sie die Analyse des modernen/kolonialen Weltsystems liefert, setzte der Niedergang der Vereinigten Staaten als Hegemonialmacht bereits um 1970 ein und beschleunigte sich während der…

Share:

Institut für die Wissenschaften vom Menschen Institute for Human Sciences

Spittelauer Lände 3
1090 Vienna/Austria
Phone +43-1-313-58-0
Fax +43-1-313-58-60
iwm@iwm.at

SPEAKER

Second African Digital Humanism Summer School, Accra, Ghana, 31 August – 05 September 2026

Francis Saa-Dittoh

Francis Saa-Dittoh is the General Chair of the Second African Digital Humanism Summer School. Senior lecturer and research fellow in ICT/AI for Development, University for Development Studies, Tamale, Ghana.

Lecture 2: *From Radio to AI – African Community-Driven Development of Sustainable Information Systems*

Reader article:

Francis Saa-Dittoh, Anna Bon, André Baart, and Gossa Lô: TIBaLLi: Internet Inclusion Through Artificial Intelligence. In L. Hagedorn, U Schmid, S. Winter, S. Woltran (Eds.): Proceedings First Interdisciplinary Science and Research Conference DIGHUM 2025, LNCS 16319, pp. 83–93. Springer Nature, Cham Switzerland, 2026. https://doi.org/10.1007/978-3-032-11108-1_6

TiBaLLi: Internet Inclusion Through Artificial Intelligence

Francis Saa-Dittoh[1(✉)], Anna Bon[2], André Baart[3], and Gossa Lô[4]

[1] University for Development Studies, Tamale, Ghana
fdittoh@uds.edu.gh
[2] Vrije Universiteit Amsterdam, Amsterdam, The Netherlands
a.bon@vu.nl
[3] Babafla, Amsterdam, The Netherlands
andre@babafla.com
[4] Amsterdam, The Netherlands

Abstract. In recent years, Artificial Intelligence has aided in the development of many solutions around the world in numerous fields and it has become imperative to ask how advanced AI methods like Machine Learning and Natural Language Processing can be reconstructed so as to make the Internet more inclusive practically, for communities in low-resource environments in the Global South. This paper outlines the project in Ghana, which utilizes a Participatory Action Research approach to build a local voice-based Automatic Speech Recognition system to provide domain-focused web-based information to local communities in Dagbani (a local Ghanaian language). We look at the methodology, and how it utilized insights from community engagement to build an inclusive system. We also look at what broader implications of this design process for the Web and AI in the context of decolonization of the Internet.

Keywords: low-resource environments · web inclusion · indigenous languages · artificial intelligence · natural language processing · participatory action research

1 Introduction

The digital divide refers to the gap between those who have ready access to computers and the internet and those who do not. This divide exists between countries, particularly the Global North-South, but also within countries (an Urban-Rural divide) which is more pronounced in the Global South and is significantly impacting social and economic development [1].

United Nations' SDG 9.c seems to suggest that the main issue concerning the Digital Divide is having universal and affordable access to the Internet. While this is indeed an important factor in the Global South, the starting point and hypothesis underlying this project is that more variables/factors are involved, in particular: (i) the local content, relevance and salience of offered Internet

L. Hagedorn et al. (Eds.): DIGHUM 2025, LNCS 16319, pp. 83–93, 2026.
https://doi.org/10.1007/978-3-032-11108-1_6

information; (ii) the local language(s) and their modality (speech; given high levels of illiteracy) in which Internet information is provided.

Decolonizing the Internet, connecting the unconnected, and closing the digital divide, does not just reside in having affordable Internet access everywhere. It is also in the ability (i) to provision locally important information (ii) in the languages and modalities people are familiar with. The project undertakes to address both variables/factors (i) and (ii). We do so by participatory field experimentations in collaboration with rural communities in northern Ghana. The project's focus is: (i) communications to and from people on highly relevant local domains such as weather/climate or market price information as pertaining to farming in the Sahel; (ii) delivering such information via local-language voice-based systems in a domain-focused speech vocabulary that is crowd-sourced from rural communities themselves.

This suggests the use of Artificial Intelligence (AI), which in recent years, has aided in the development of many solutions around the world in numerous fields. It has therefore become imperative to investigate how AI can make impact in resource-constrained environments. In this paper, we therefore look at an innovative and practical example of how AI can be leveraged for digital information access in resource-constrained environments and explain why the concept of stakeholder and community participation is imperative right from the ideating phase.

2 Methodology

The research adapts a Participatory Action Research (PAR) methodology where individuals and communities are conceived of as first-class partners in a co-creation. This is in contrast to certain social science approaches whereby people (whether called objects, subjects, or even research participants) are in fact considered as "observed entities" by a value-neutral academic "spectator" researcher, as for example is the case in empiricist variable-oriented hypothesis-testing [2,3]. A general overview of PAR is found, e.g., in [4]. As a methodological paradigm, PAR has a long history and tradition of emancipatory aims and roots [5,6] and so fits well into critical indigenous [7] and decolonizing methodologies [8,9]. With respect to the design and deployment, we employ Design Science methodology for Information Systems and Software Engineering as explicated in [10] that is concerned with constructing, testing and evaluating ICT/IS artefacts in interaction with their real-world context. This methodology is central and key to the research, as failures in ICT implementations in the Global South have shown that a human-centered approach is not only helpful but imperative to successful design, adoption and use [11–13].

In this research, we begin this process by way of stakeholder meetings with community members, community leaders, farmers, agricultural extension workers, agricultural researchers, climate experts, linguistics experts, AI experts, etc. The various meetings and engagements led to a stakeholder workshop where, with the end-users (farmers from the targeted communities) as the main focus,

we delved into the various aspects of the research question. Day 1 of the workshop had participants formed into break-out sessions consisting of a combination of expertise. A typical group contained about 9 members including 2 farmers (one male, one female), a moderator and scribe (mostly a core team member), and about 5 others ranging from ICT4D Researchers, AI Experts, Linguists, etc. Day 2 had a different structure; groups were structured based on expertise; 2 AI Groups (AI Experts, AI Enthusiasts, Programmers,....), 1 Language Group (Linguistic Experts, Media Experts, Native Dagbanli Speakers,...), 2 Climate and Farmer Group (Meteo Experts, Agriculture Experts, Farmers,....).

This further led to a larger community engagement at the communities, where the team held focus group discussions (FGDs) with almost 100 community members in three separate groups; Men, Women and Youth. These focus group discussions centred on the same concepts from the workshop; ICTs, Farming, Climate and Language, but broken down for the understanding of a wider audience for discussion. For the most part, these discussions reiterated the accuracy of the information that was teased out during the sessions. In addition to the knowledge gleaned for the continuation of the project, the FGDs were recorded and kept for future use.

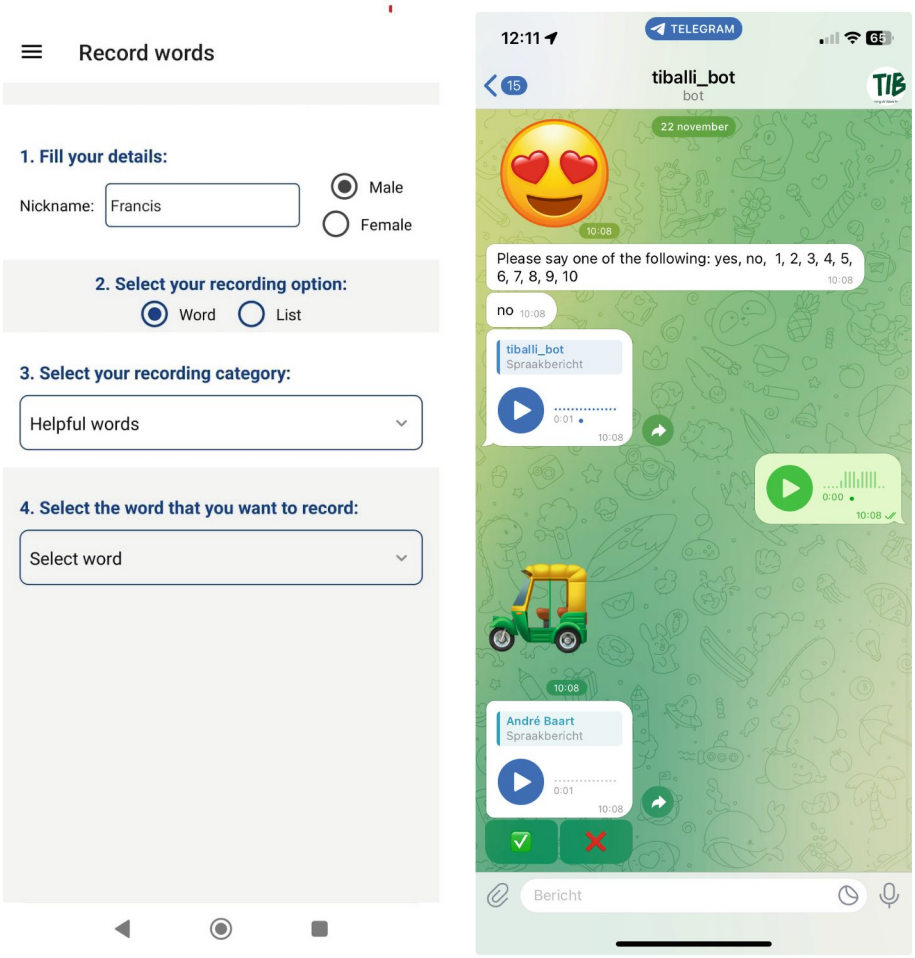
The result of these in-depth deliberations was a detailed road-map resulting from the research question, but most importantly, in context of the local situation and centered on the end-users and stakeholders.

3 Implementation and Intermediary Results

To realize the project's goal of implementing a domain-focused Automatic Speech Recognition (ASR) model for a small indigenous language (Dagbani, spoken by approximately 3 Million people in the Northern Region of Ghana), it was required to 1) obtain recorded fragments for words in the selected domain (we began with numbers 0 to 10, "yes" and "no"), 2) using Machine Learning, build a Natural Language Processing (NLP) model for our dataset, 3) incorporate our NLP model into a system that is usable by the local community.

For the first step, team members worked together with people within the target communities to build a crowdsourcing app[1] that permits users to record people pronouncing local words. These recordings will serve as training data for the machine learning model that will accurately convert Dagbani speech to text. The mobile application is developed with React Native and Expo. It offers an intuitive interface for capturing and submitting audio samples of local words. Utilizing the capabilities of mobile devices, the application ensures high-quality recordings. The app allows for the recording of voice fragments. A list of words are provided (currently; "yes", "no" and the numbers 1 to 10) which the user can select and then prompt a willing participant to record said word in a local language. The recorded fragments are stored (temporarily) on the device and can be later uploaded to our server (Firebase). Figure 1a shows a screenshot of the App.

[1] Source code will be linked.



(a) Data Collection App (b) Telegram Bot Prototype

Fig. 1. Screenshots of the tools developed.

This simple, but effective tool is used by enumerators to take recordings from community members. During the first iteration of the model, almost 4000 voice fragments were gathered in roughly 4 months with only a small data-collection team of about 8 users. The process of building this app was not trivial, especially in terms of PAR; ensuring that the end-users and stakeholders are involved in the process so as to iteratively build a solution that really works both technically and contextually in the local setting. The concept of a localized app, built with contributions for local communities and used by actual community members for crowd-sourcing in itself is quite novel in the context of resource-constrained communities and came about due to the user-centric, community-

based, participatory mindset resulting from the methodologies learnt and used over the years [14].

AI Model Build Process. With the gathered training data of approximately 4000 utterances of Dagbani words, an AI model was trained. The base model used was Wav2Vec Massively Multilingual Speech, by Meta [15]. This is a model based on the well-known wav2vec2 architecture that is pre-trained on 1100 languages. We fine-tuned this model, using the training data of our limited initial vocabulary (yes, no, 0 through 10). This resulted in a model that is tuned to these specific words and will only give back results that fall in this vocabulary.

Telegram Bot and Test. In order to test the model, we built a Telegram bot that 'parrots back' the user. We chose Telegram as it is well suited for rapid development of prototypes. The prototype bot expects a voice message from the user, containing one of the words in the vocabulary. This message is then processed by the ASR model, which results in a written version of what the user said. This text is then converted back to speech (using the slot-and-filler method [16]) and sent as a voice-message back to the user. The user can then listen to the response of the bot, and confirm whether the result was correct or not. We included two rapid-response buttons, a green checkmark (correct) and a red cross (incorrect), which the user can use to indicate the quality of the result. The result is stored and can be used (together with the user input) as additional training data to improve the accuracy of the model in a later stage. Figure 1b shows a screenshot of the Telegram Bot.

The prototype was tested, as seen in Fig. 2 in the Tingoli community, which resulted in a near-perfect accuracy through the diverse population (young, old, male, female, etc.).

4 Lessons Learned and Reflections

The project's successes, so far (successful corpus collection, development of AI Model, implementation into a test system, testing in the field), shows us that the question of advanced AI methods (ML, NLP) being reconstructed so as to make the Internet more inclusive for communities in low-resource environments in the Global South, is doable, especially in a considerably short amount of time. The high accuracy of the AI model shows that it is not required to have an overly large dataset for successful ASR in small languages.

4.1 Insights from Community Engagement

The methodology employed in this research requires collaboration with many entities which all should focus on the community. Not surprisingly, there is little to no opposing needs or priorities shown within our select communities, nor between communities and stakeholders. This is because collaboration with these various stakeholders (farmers, agricultural institutes, Universities, Ministry of Agriculture, linguists, etc.) did not by any means start with us. These groups

Fig. 2. Prototype under testing in the community.

(although perhaps not the same communities and individuals) have worked hand in hand for a long time in developmental issues especially pertaining to agriculture. Due to this, their challenges and needs have been clear to them over time.

This existing relationship enabled easier community engagement and collaboration between all involved entities. This led to very insightful information from the communities which aided the ongoing successes of this project. The nuances of dialects (which as "outsiders", we may have overlooked), the complexities of numbering (some numbers - e.g. 1 - have more than one word), the complexities of currencies (numbering for currencies is completely different, and more complicated, than counting), the traditional nuance in the non-word "Zero", which does not exist as a number, are some examples of intricacies that community engagement did not just bring to light, but helped to posit solutions for.

Beyond this, community members played a major role in iterative feedback for the design and build of the data collection app, select members were trained and handled the data collection process in the field, and large groups of community members participated in testing the ASR implementation.

We see from this experience (in addition to lessons learnt from numerous failed and successful ICT4D implementations) that the involvement of the intended users of solutions for low-resource environments is imperative. Regardless of the methodology used, there should be constant community and stakeholder involvement; remember, we are "plugging in" our innovation into their ecosystem to solve a problem they have identified together with us. The lifecycle of the solution development must therefore include end-user and stakeholder

engagement, with prototyping where they can utilize the modules of your solution to see if they fit in with what they require and are inclined to use.

4.2 Inclusivity

The project was especially particular in ensuring the collective voice (literally and figuratively) of men, women and youth. Community engagement was structured to ensure that all opinions were heard. Moreover, the success of the ASR model required voice fragments from all genders and diverse age groups. For the most part, the opinions, challenges and suggested solutions were similar across genders and age groups. However, there are, of course, some areas of interest pertaining especially to men, women and youth respectively; while men and youth are mostly concerned with staple and cash crops (maize, rice, yam, etc.), women are more concerned with vegetables (pepper, tomatoes, green leafy vegetables, etc.) which tend to have different informational needs in terms of farming practices, market prices and climate. This, luckily, is not a huge challenge in this project, since there is available information on for all these crops and climate information provided can cater for all crops. It must however be ensured that any system built, for agriculture, is not biased towards crops that are of interest to a particular gender or age group, especially since at the onset, a few specific crops may be selected for the system.

4.3 Participant Benefits, Skills, and Ownership

The PAR-based methodology of the project enabled local data gathering, which involved local enumerators with financial remuneration to go out, record short voice bits, check the quality and write down what participants say. They also benefit from learning how to use a phone for recording, how to tag information, and how to fix simple technical problems. In the workshops, the volunteers saw how to run the little audio tool themselves and check the recordings. At the same time, students from the University for Development Studies gain from integration of aspects of the project to build and test the tools in their assignments and mini projects. This ensures that some technical skills are passed on after the pilot is over. Finally, the open Dagbani dataset [17] lists the Tingoli and Nyankpala communities as coauthors (with permission), operationalizing community credit and citation in perpetuity; this is a form of ownership that remains with the corpus and any future reuse.

4.4 Sustainability

Keeping the project alive requires several pieces. First is the technical aspect, which is sometimes overlooked. However, technical choices can have a huge impact on sustainability. The system runs on GSM-based IVR instead of heavy data apps, so it works on cheap phones and needs almost no Internet. The next step will use a small plug computer that does phonetic matching and a tiny

speech recognizer, hosted at a local place like UDS/ISOC Ghana and looked after by Ghanaian IT staff, which cuts the need for big cloud services. Second, the partnership with UDS, ISOC Ghana, and SARI gives the project a solid home for hardware, data, and model files once the pilot is over. That makes it credible and gives logistical help. Third, the financial side remains stable by keeping the paid recorder jobs and turning some of them into local AI-tech apprenticeships, so people move from just collecting data to also testing and analyzing it. The whole tool chain – a tiny ReactNative plus Firebase app, a Docker container for speech matching, and simple bot fronts – is rolled out at the start of the rainy season, which helps repeat the work in new places.

4.5 Technical Challenges and Solutions

Our biggest technical challenge interestingly is not from Artificial Intelligence, as one may expect in this attempt to use advanced technologies in low-resource environments, but as a result of the initial information type selected for the use-case; climate. During the research, it was revealed that due to the lack of sufficient ground weather stations in Ghana, there is extremely low accuracy in weather prediction and as such it is not trusted by local farmers [18]. This meant that the information being delivered, if we attempted climate information in its current state, would be inaccurate and would result in mistrust of the system. To mitigate this, and also to continue the main aim of the project; using advanced AI methods (ML, NLP) to make the Internet more inclusive for communities in low-resource environments, we did two things; 1) launch a sub-project on improving climate information by resourcing local weather information and 2) focus on readily available and more accurate, but also very relevant information; market prices.

4.6 Broader Implications for Web and AI

The same principles that have governed the successful ICT solutions in low-resource environments that provide information from the web (example in [19]) thus far are largely the same principles required for successful use of AI for solutions in low-resource environments, but even to a much larger extent. The fathers of the Internet have at recent times mentioned how colonized their invention has become and how on hindsight it would have been better to have certain voices at the start, which was not the case. The same "mistake" should not be repeated in the AI revolution, especially in solutions for marginalized communities.

The successes of this project thus far shows us that community involvement is as paramount as the technical skills needed to build these solutions. More so, it is imperative that the technical designers of these solutions have a good understanding of the local context, which comes from the communities and stakeholders, inherently forming an inter-disciplinary team albeit not necessarily an academic one. This leads to a human system, literally built by the people, for the people.

Lastly, the idea of a participatory design process focused in the field directly promotes decolonization. In this project, the idea is local, the sourced voice data is local, the voice data collection process is local, the NLP model is based on this local data, testing and evaluations are done locally in the field. The data to be shared may be local or not, but again, this is based on the needs and wants of the communities. This already combats issues like bias in our AI model. There is of course still reliance on external servers and LLM architecture which in time can also be localized. We wish to gradually shift the dependence on large technology companies on the road to decolonization [20].

5 Future Work

The project will build a larger corpus on numbering, currencies and crops. This will follow the same methodology and will result in a local-based ASR system that community members can interact with to obtain web-based market prices of crops from different parts of the country. Future directions will also involve a look at alternative methods of building ASR. One promising direction is the concept of phonetic-based ASR methods which will greatly reduce the need for large datasets for building ASRs for small languages.

Acknowledgments. This work was supported by the Internet Society Foundation (https://www.isocfoundation.org/), the CSIR-Savanna Agriculture Research Institute, Ghana (https://csir-sari.org/) and the Web Alliance for Regreening in Africa (W4RA) (https://w4ra.org/), with deep gratitude to Wilhelm Nomu Kutah, Gideon Amakama Ali, Bram Kreuger, Chris van Aart, Hans Akkermans, members of the Internet Society Ghana Chapter (https://isoc.gh/), and the people of Tingoli and Nyankpala in the Tolon District of Ghana for their enthusiastic participation and insights.

References

1. Ragnedda, M., Muschert, G.: The Digital Divide: The Internet and Social Inequality in International Perspective (2013)
2. Robson, C., McCartan, K.: Real World Research, 4th edn. Wiley, Chichester, UK (2016)
3. Johnson, C.W., Parry, D.C.: Fostering Social Justice Through Qualitative Inquiry. Taylor & Francis Group (2022)
4. Reason, P., Bradbury, H.: Handbook of Action Research. Sage Publications Ltd. (2006)
5. Freire, P.: Pedagogy of the Oppressed. Continuum (2000)
6. Borda, O.F.: Investigating reality in order to transform it: the Colombian experience. Dialectical Anthropol. **4**(1), 33–55 (1979). http://www.jstor.org/stable/29789952
7. Denzin, N.K.: Handbook of Critical and Indigenous Methodologies. SAGE (2008)

8. Kovach, M.: Indigenous Methodologies: Characteristics, Conversations, and Contexts. University of Toronto Press (2010)
9. Smith, L.T.: Decolonizing Methodologies: Research and Indigenous Peoples (2012)
10. Wieringa, M.: What to account for when accounting for algorithms: a systematic literature review on algorithmic accountability. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (2020). https://doi.org/10.1145/3351095.3372833
11. Benjamin, P.: Community development and democratisation through information technology. In: Reinventing Government in the Information Age. Routledge (1999)
12. Tettey, W.: The media and democratization in Africa: contributions, constraints and concerns of the private press. Media Cult. Soc. **23**(1), 5–31 (2000)
13. Sharma, V.V., Rao, K.V.: Bridging the digital divide: information kiosks in rural India. In: Rural India: Achieving Millennium Development Goals and Grassroots Development, vol. 383 (2009)
14. Panayiotou, A., Saa-Dittoh, F., Bon, A.: Resourcing small indigenous languages in the field – designing a user-centered data collection method for automatic speech recognition. In: Proceedings of ICT4SD 2023, vol. 3, Goa, India (2023)
15. Pratap, V., et al.: Scaling speech technology to 1,000+ languages. arXiv:2305.13516 [cs.CL] (2023)
16. Baart, A., Bon, A., de Boer, V., Dittoh, F., Tuijp, W., Akkermans, H.: Affordable voice services to bridge the digital divide: presenting the Kasadaka platform. In: Escalona, M.J., Domínguez Mayo, F., Majchrzak, T.A., Monfort, V. (eds.) WEBIST 2018. LNBIP, vol. 372, pp. 195–220. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-35330-8_10
17. Francis Saa-Dittoh and Tingoli and Nyankpala Communities of the Northern Region of Ghana. Recordings of spoken Dagbani for the letters zero to ten and the words yes and no. Zenodo (2024). https://doi.org/10.5281/zenodo.13284004
18. Nyadzi, E., Werners, S.E., Biesbroek, R., Ludwig, F.: Towards weather and climate services that integrate indigenous and scientific forecasts in Ghana. Environ. Dev. **42**, 100698 (2022). https://doi.org/10.1016/j.envdev.2021.100698
19. Dittoh, F., Akkermans, H., de Boer, V., Bon, A., Tuyp, W., Baart, A.: Tibaηsim: Information access for low-resource environments. In: Yang, X.-S., Sherratt, S., Dey, N., Joshi, A. (eds.) Proceedings of Sixth International Congress on Information and Communication Technology. LNNS, vol. 235, pp. 675–683. Springer, Singapore (2022). https://doi.org/10.1007/978-981-16-2377-6_62
20. Wieringa, M.: What to account for when accounting for algorithms: a systematic literature review on algorithmic accountability. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (2020). https://doi.org/10.1145/3442188.3445885

Digital Humanism Initiative
Technology for Humanity

SPEAKER

Second African Digital Humanism Summer School, Accra, Ghana, 31 August – 05 September 2026

Anna Bon

Anna Bon is researcher and lecturer in Information and Communication Technologies for Development (ICT4D) at Vrije Universiteit Amsterdam, and senior project manager at VU Amsterdam's International Office (CIS Centre for International Cooperation). She directs the interdisciplinary research, education and community service program W4RA (Web alliance for Regreening in Africa). Her interdisciplinary research centers around how innovative context-aware technologies, serving especially communities in Africa's drylands. Her international projects include digital voice-service support for local value networks in West Africa; ICT4D Community Service Learning in rural Sarawak, Malaysia; Sustainability and Ethics in Digital Development; AI for Sustainable Development, and the Amsterdam Digital Divide.

Lecture 3: *Africa and Digital Humanism* (Round-table plenary discussion)

Reader article:

Anna Bon, Francis Saa-Dittoh, and Hans Akkermans: Bridging the Digital Divide.
In H. Werthner, C. Ghezzi, J. Kramer, J. Nida-Rümelin, B. Nuseibeh, E. Prem, A. Stanger (Eds.): Introduction to Digital Humanism – A Textbook, pp. 283-398. Springer Nature, Cham, Switzerland, 2024. https://doi.org/10.1007/978-3-031-45304-5_19

Bridging the Digital Divide

Anna Bon, Francis Saa-Dittoh, and Hans Akkermans

Abstract This chapter discusses the disparity in access to digital services, which exists between countries, regions, communities, and people in the world. This disparity is referred to as the *digital divide.* Digital information and communication are obviously of key importance for the development of countries and regions. However, different approaches exist to address this problem. In this chapter, from a digital humanist perspective, we explore practice-oriented digital design approaches to serve people and communities in currently unconnected regions of the world. We discuss how this might best be done, upholding ethical standards, inclusivity, and human-centered principles. To illustrate this in action, we present a case study from rural Ghana.

1 Introduction

The digital divide refers to the disparity between countries, regions, and people in their access to digital services (Fuchs & Horak, 2008; Van Dijk, 2020; Potter et al., 2008; Mubarak, 2015; Bon, 2020). For example, in 2023, 57% of the African population is unconnected to the global digital society (Internet World Stats, 2023). This disparity is considered a global challenge. It is obvious that digital information and communication are of key importance for the development of countries and regions.

A. Bon (✉)
Centre for International Cooperation, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands
e-mail: a.bon@vu.nl

F. Saa-Dittoh
University for Development Studies, Tamale, Ghana

H. Akkermans
AKMC, Koedijk, The Netherlands
e-mail: hans.akkermans@akmc.nl

H. Werthner et al. (eds.), *Introduction to Digital Humanism*,
https://doi.org/10.1007/978-3-031-45304-5_19

Different approaches exist to address the digital divide. Some approaches are technology-focused, seeking progress through technological advancement. Some approaches focus mainly on profits and economic growth. Other approaches are more human-centric, trying to achieve human well-being. Each approach reflects a different worldview or school of thought, with varying assumptions about the nature of development, the role of technology, and the desired outcomes of bridging the digital divide.

From a technology-focused perspective, *not being connected* is attributed to a lack of physical and digital infrastructure. Due to the absence of proper infrastructures in poor regions or countries, many people do not have access to digital services and content. This withholds their participation in the information society and hampers development. Rolling out digital infrastructure and making the Internet available to people in every corner of the planet is then the solution. The ensuing question is how to organize this.

From a slightly different perspective, the digital divide is only a snapshot in time, in an ongoing process of technological innovation and diffusion (Rogers, 2003). Poor regions are lagging but will soon catch up in their technological and organizational capacities and will adopt digital technologies at a later stage. This will lead to progress and bring economic development, at long last (De Janvry & Sadoulet, 2021; Taylor & Lybbert, 2020). The uptake of digital technologies by people in low-resource environments is seen as an opportunity, as this will create a new market segment for technology vendors and other markets. This market segment is referred to as the "bottom of the pyramid" (Prahalad, 2004; Heeks, 2008). Economic growth is the main justification for this approach.

Both perspectives—technology-focused or economy-focused—have one idea in common: digital technology is the motor for progress and development. Access to digital information and services will lead, from an "underdeveloped" situation to a situation of prosperity, like in the industrialized countries (Wicander, 2009). It is assumed that well-being will follow from technological and economic development and that prosperity will *trickle down*, naturally, to most people.

Alternative perspectives on the digital divide value a more human-centered standpoint. Digital humanism, decolonial theory, and grassroots initiatives are examples of human-centered approaches. They prioritize human well-being, social inclusion, equity, and emancipation above technology and profit.

In 2019, the *Digital Humanism Initiative* was launched in Vienna by a group of primarily European scientists, policy makers, and practitioners concerned about the disrupting impact of digital progress on society, economy, and the natural environment. The Vienna Manifesto for Digital Humanism was proclaimed as a call for ethics and humanism and as an antidote to technocratic approaches with their focus on merely technological innovation and economic growth (DIGHUM, 2019). The digital humanism movement is now supported by many scholars, politicians, and policy makers from around the world. It launched a debate about the purpose and the future of the digital society (DIGHUM, 2019). At the onset, the Digital Humanism Initiative focused on the societal issues of the digital society in the Global North, where it originated, but soon it embraced a cosmopolitan viewpoint, in which the

Global South and the exclusion of people in low-resource environments are a central problem to be discussed and addressed.

A similar, and even more radical, perspective is found in decolonial studies, in which patterns of inequality, poverty and wealth, domination, and colonialism are explored and analyzed in society (Mignolo & Walsh, 2018; Mendoza, 2021). Decolonial theory offers an alternative lens to assess the digital divide and explore its root causes and origins, in the light of colonialism. The conjecture here is that the digital divide is inherently unfair because it excludes people. However, the idea also exists that digital technologies, if well designed, may be able to help improve certain aspects and overcome certain problems of people. Of course, these digital technologies must be, by design, respectful of human rights and human dignity and of local agency, culture, norms, and values.

Whereas academic studies, e.g., Lin et al. (2015) and De et al. (2018), express criticism about various aspects of the digital divide, only a few of these studies have been action oriented. Yet, from a digital humanism perspective, the purpose and goal are not only to discuss but also to influence policy, raise awareness, and call for practical, problem-solving action. An example could be an interdisciplinary study that combines action research, design science, and a human-centered approach to design technologies that support people and communities in low-resource environments. The persistent question is how this study can ensure fairness, ethics, and inclusion by design while at the same time being respectful at the autonomy and agency of local users, communities, and people.

This question will be discussed in the sections that follow. In Sect. 2, we give a brief history of the digital divide. In Sect. 3, we observe the digital divide through a decolonial lens, assessing roots, causes, and patterns of inequality. In Sect. 4, we seek the requirements for inclusive, human-centered design. In Sect. 5, we propose a human-centered methodology that combines action research, design science, and ethics perspective. We illustrate this through a case study from northern Ghana in Sect. 6.

2 Short History of the Efforts to Bridge the Digital Divide

Since the 1990s, many efforts have been undertaken by the international community[1] including international organizations such as the World Bank, the United Nations, UNESCO, national and international development agencies, numerous non-governmental organizations (NGOs), and large private charity organizations to make digital technologies available for social and economic development of poor regions of the world. Despite large budgets and numerous projects in prestigious international development programs, a mismatch between technologies and the

[1] We refer to these entities as the "international community." This term is used in international relations to refer to a group of people, organizations, and governments in the Western world.

target environment often occurred (Bon, 2020). Many unsuccessful technology transfers resulted from blind optimism about the impact and reach of digital technologies, combined with poor understanding of local needs and context. Techno-optimism about digital technology culminated in the United Nations World Summit on the Information Society, in 2005, in Tunis, where heads of states from all around the world and global players in the big private technology sector came together, including Microsoft, HP, Nokia, and Intel (Berry, 2006; Qureshi, 2006). Goals were set during this summit, to roll out the Internet before the year 2020, in every corner of the planet.

Soon after 2005, it became clear that many ambitious, large-scale development projects were failing, especially in serving the poorest and least connected communities with digital technologies (Bon, 2020). Two examples are (i) the *One Laptop Per Child* (OLPC) project, which started around 2008 and aimed to provide every schoolchild in the world with a laptop (Buchele & Owusu-Aning, 2007), and (ii) the Millennium Villages project, which started in 2005 and selected a small number of very poor and unconnected villages across African countries to pull them "out of poverty" by giving them an investment boost, transferring state-of-the-art Western technology, and connecting the communities to high-speed Internet (Sanchez et al., 2007). Both the OLPC and the Millennium Villages project spent large sums but did not meet their objectives or expectations[2]. The failures of these and many other large-scale technology transfer projects tempered the initial techno-optimism of the international community.

Unexpectedly, it was not the Internet, but simple, mobile, voice-based telephony, rolled out by the telecom companies in many remote and poor regions to low-income communities of the world, that became a huge success and had a large impact for people's lives at the so-called base of the pyramid (Kalba, 2008). For many poor people in urban or remote rural areas in developing countries, mobile telephony was their first digital experience and remote communication. The mobile telephony success was followed by an innovation called *mobile money*. This technological innovation, which only required a simple mobile phone—no smartphone or Internet connection—enabled people to deposit money on their cell phones and transfer it to other users of the service, without need for a bank account. Mobile money was first launched in 2007 in Kenya, as a truly African innovation by the local telecom operator Safaricom. This service registered 1.1 million users in the first 8 months after its inception (Mbiti & Weil, 2011). Soon, many countries in Africa followed, as Orange, Airtel, and many other telecom companies started providing mobile payment services to millions of previously unbanked citizens in Africa (Nan, 2019).

So, in the past two decades, the least developed countries in, e.g., sub-Saharan Africa have caught up very rapidly and entered the digital era. However, innovation and technology adoption often followed a different path than in the Global North. For example, many people in rural Africa skipped text-based systems and are now

[2] A full discussion about these and other cases of unsuccessful ICT4D interventions is given by Bon (2020).

using asynchronous voice-based communication through voice messaging systems and social media apps, without ever having sent a written (text-based) email (Dittoh et al., 2020a, 2020b).

Despite the positive impact of mobile telephony and mobile money for many people in rural Africa, the profits remained unequally shared. At the onset of the 2020s, many sub-Saharan countries were mainly *consuming* digital services but hardly *developing* or *producing* any.

From the point of view of infrastructure, only few data centers or Internet exchange hubs exist in sub-Saharan countries (Augustine, 2022). The data produced by African users are stored in the cloud, i.e., remote data centers[3]. The costs for data transport between users and the "cloud" make digital services more expensive in remote areas than, e.g., in urban centers. This is another example of inequality.

Technological innovation is moving fast. The technology leap in the second decade of the twenty-first century is based on data-driven artificial intelligent systems. Many new beneficial solutions are to be expected, e.g., for the development of speech recognition of local indigenous languages or for the digitalization and access to specific, contextual knowledge systems. Yet, big concerns exist, related to the deployment of generative AI.

People in countries in the Global South may be vulnerable to the spread of disinformation and digital surveillance. They may face infringement of their intellectual property rights, due to lack of legislation and governance of the digital sphere (DIGHUM, 2019). Since privacy rules and regulations such as GDPR are still not implemented in many African countries, there is no good protection of people's identities, privacy, and data ownership. Especially people "at the base of the pyramid" may be less well informed and therefore more vulnerable to predatory technology firms and digital scams. Moreover, in the light of a rapidly involving technology like artificial intelligence, combined with a lack of policy and legislation in many countries in the Global South, this may result in a complete governance of the digital society, including the infrastructure, data, and technological know-how, to be concentrated in just a few current market-dominant private technological firms, also known as Tech Giants, such as Meta (Facebook), Alphabet (Google), Amazon, Apple, and Microsoft.

[3] To reduce costs of data transport, digital data is usually stored as close as possible to its consumers. In the absence of data centers in many African countries, the nearest locations are data centers in Europe and the Middle East. Upcoming private investments in the richer African countries, including South Africa, may bring new data centers to the continent.

3 The Digital Divide Through a Decolonial Lens

The digital world, where we socialize, interact, and take decisions, mirrors the physical world in many aspects. This includes patterns of inequality and domination, which are often referred to as "coloniality" (Mignolo & Walsh, 2018; Mendoza, 2021 pp. 46–54; pp. 1–12; Quijano, 2007 pp. 15–18; Hope, 2022). Decolonial theory helps uncover hidden patterns of domination within social structures. In this discussion, we explore a few examples from the digital society.

One common assumption, also found in the SDG9, is that many underdevelopment-related problems, e.g., in education and trade, will be solved, once Internet connections have been established. Yet is this a valid assumption? In the light of the concerns about artificial intelligence and the increasing concentration of knowledge and power, it is debatable whether the Internet will bring democracy, social justice, equality, and a sustainable and prosperous life to all people (Bon et al., 2022).

Let's take the example of "free Internet", offered by big tech firms, such as Google (Dahir, 2020), Starlink (Sapah, 2023), and Facebook (The Guardian, 2016), to people in low-resource environments. These initiatives will ultimately connect many people, even in remote low-resource environments, to the digital backbones for free. However, we must question what "free Internet" means in a capitalist world. The revenues of the large technology companies rely on extracting value from personal data, people's Web browsing preferences, by tracking their entire online behavior using AI algorithms. Users are often unaware of the exploitation of their personal data. This business model of extracting value from the people's online behavior has been criticized by Shoshana Zuboff in her book *The Age of Surveillance Capitalism* (2019), for which she introduced the term "behavior surplus."

Another aspect of decolonial critique is the hegemony of the Internet as a network standard. While the Internet can be seen as a global common, a platform to share information, access opportunities, and collaborate across geographic and cultural boundaries, as discussed earlier, it has disadvantages. Choosing not to be part of this network results in isolation, for the non-user. This characteristic of standards and networks, which is described by Grewal (2008, pp. 20–28), makes the Internet a hegemonic system without escape, despite the price users, communities, and even countries pay with money or data, to access it.

Geolocation and governance of digital infrastructures are also aspects related to coloniality. The digital society is heavily concentrated, physically, economically, and socially, in the Global North. The commercialization of digital technologies, influenced by this centralization, further channels the profits from innovations in the Global South to investors (Zuboff, 2019, pp. 63–96). Governance and decision-making in this realm are predominantly controlled by private tech firms, operating under norms and regulations from countries in the Global North. Unequal competition in terms of storage, connectivity, funding, and adoption also hinders start-ups in the Global South (Bon et al., 2022).

In many countries, big tech firms and telecom providers are assuming public roles and functions traditionally held by the state, such as in healthcare, education, and infrastructure. Through philanthropic gifts and corporate social responsibility, digital services are provided in exchange for data, market penetration, tax savings, branding, and policy influence. Youngsters are often targeted by big tech companies with media, music, video, entertainment, and news (Pini, 2020, pp. 37–40).

Coloniality is also evident in artificial intelligence (AI). Machine learning algorithms, previously assumed to be objective and value-free, have been found to harbor many discriminating biases (Mohamed et al., 2020, pp. 659–663). These biases are often concealed in the underlying data. Trivial examples include visualization programs that autonomously lighten the complexion of black and Asian faces or smart doors that, based on facial recognition algorithms, only open for white faces while remaining closed for a person with dark skin, as the machine learning algorithm fails to recognize them as a person. These racial biases can emerge unexpectedly in autonomous smart systems. If these algorithmic biases are not properly addressed, they will further perpetuate inequalities and injustice against certain groups and individuals (Mohamed et al., 2020, pp. 659–663).

The aforementioned discussion highlights the concerns regarding coloniality, domination, biases, and injustice. If left unaddressed, these issues will result in a digital society that is exclusive, unethical, and, from various perspectives, reminiscent of neo-colonialism.

It is crucial to recognize that innovation in digital technologies is not an autonomous process, but one driven by deliberate choices. Methodologies from information systems engineering provide flexible and powerful approaches for designing and fostering open innovation. Action research with its long-standing tradition of improving real-world situations adds to it the human-centered and ethical aspects. In the following section, we will delve into the discussion of how information and communication technologies can be designed in a human-centered manner.

4 Requirements for a Human-Centered Approach

When exploring alternative approaches to serve communities and individuals in low-resource environments, which principles are required?

The foremost and crucial principle is that digital technologies should not cause harm to anyone. It is imperative that digital services align with the objectives and goals defined by the local people and users. To achieve this, local users must actively participate in decision-making processes regarding the goals and objectives of the digital service.

Secondly, digital technology must be adaptable in a flexible way to the local context, enabling it to overcome local barriers. These barriers can include language, literacy, limited infrastructure availability, local purchasing power, or other context-related issues.

Thirdly, there must be local ownership of data and the protection of local domain knowledge and intellectual property. This is essential to prevent compromise or misuse of these valuable assets. In the case of co-design, ownership of the designed artifacts should also be shared. Research data should not be automatically owned by the researcher or published without permission or consent.

In summary, the development of digital services for unconnected people or communities in low-resource environments should be conducted in collaboration with local users. Since software developers and engineers often lack familiarity with the low-resource locations in question, significant emphasis is placed on communication between users and developers through collaborative workshops and co-design sessions.

In the next section, we will describe a methodology and approach that was developed in the field from 2009 to 2019 through an action research program named W4RA—the Web Alliance for Regreening in Africa (W4RA, 2023). The program's objective is to support local smallholder farmers in three West African countries, Mali, Burkina Faso, and Ghana, by designing digital technologies to help them achieve their objectives. The approach is referred to as Information and Communication Technologies for Development, version *3.0*, abbreviated as ICT4D 3.0 (Bon, 2016).

5 Combining Action Research, Design Science, and Ethics Perspective

ICT4D 3.0 represents a grassroots approach to bridging the digital divide, offering an alternative to the prevailing, economic growth-oriented approach to bridging the digital divide. This incumbent approach is described by Richard Heeks (2008) as "a new opportunity for ICT vendors to harness digital innovation and serve the world's poor, profitably," in an article, titled "ICT4D 2.0: The Next Phase of Applying ICT for International Development."

ICT4D 2.0 was presented by Heeks as a great improvement to the previous approaches, in which off-the-shelf digital technologies were transferred to poor regions, often with limited success (2008, p. 26).

In contrast to its predecessors, ICT4D 3.0 positions itself as a human-centered, grassroots methodology for designing digital technologies based on users' needs. It follows a five-step framework:

(i) Context analysis—understanding the users' environment
(ii) Needs assessment—understanding the user's goal
(iii) Use case and requirements analysis—defining and specifying a feasible digital solution
(iv) Prototyping, engineering, and deployment of a digital service
(v) Sustainability analysis—exploring the local business ecosystem to ensure the long-term availability of the service

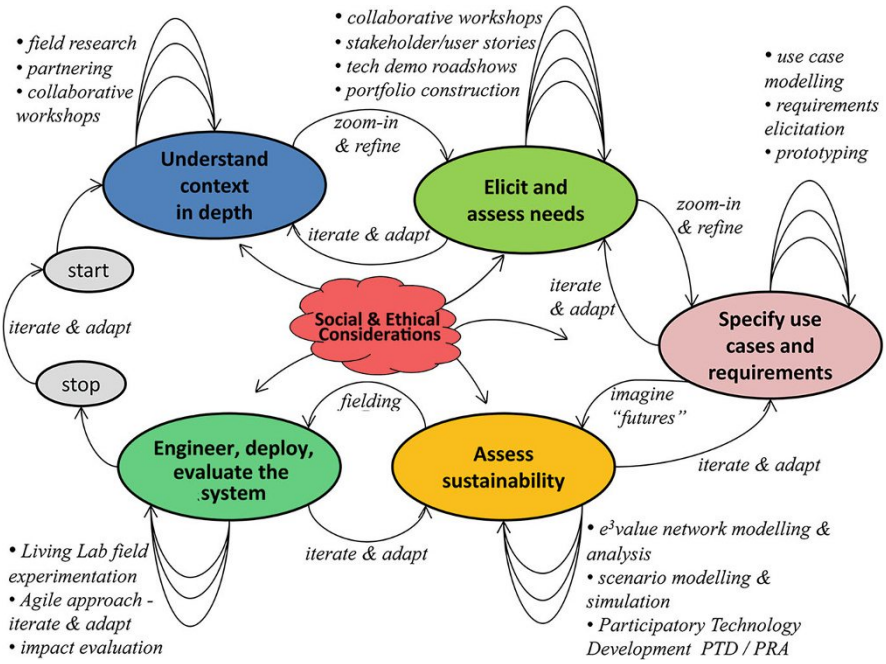


Fig. 1 ICT4D 3.0 as a collaborative, iterative, adaptive approach to digital development in low-resource environments. From: Bon, A. (2020, p. 204, *Intervention or Collaboration—Redesigning Information and Communication Technologies for Development*, Amsterdam: Pangea)

ICT4D 3.0 is a goal-oriented approach, striving for a certain improvement for its envisaged users. A conceptual visualization of this approach can be represented as a goal-strategy map (Fig. 1, cf. (Rolland, 2007)), where each colored ellipse signifies an intention to be achieved through corresponding actions. The process is iterative and adaptable, allowing for adjustments based on new information.

Emancipation, autonomy, and inclusion are the core values ICT4D 3.0 strives to achieve. It has undergone continuous evaluation and validation by its users in the field (Bon, 2020). Importantly, it can be considered a decolonial approach as it empowers users to make decisions rather than having external parties impose solutions upon them.

6 Tiballi: A Case Study of AI and Data Science for Farmers in Ghana

A case study that illustrates the ICT4D 3.0 approach in a community-oriented, transdisciplinary research project is *Tiballi*, a research project, set up in 2023 in northern Ghana (Tiballi, 2023; Dittoh et al., 2021). Tiballi explores how artificial

Fig. 2 Tiballi project's needs assessment, during a focus group discussion in the village of Tingoli, community of Nyankpala, Ghana. Photograph by Gideon Amakama Ali (18 February 2023)

intelligence—and more specifically machine learning and natural language processing—can serve people in a low-resource environment. Tiballi's envisaged users are proficient in a local, under-resourced African language only; they have low incomes, low literacy skills, and limited access to the Internet.

Tiballi's project goal aligns with an explicit wish of farmers in northern Ghana to have access to rainfall data, as this information is essential in rain-fed agriculture, but often not accessible. These wishes were collected during workshops and focus group discussions; see Fig. 2. In these meetings, it becomes clear that (a lack of) Internet access is not the only barrier for the community to use digital services. It is also important to provision locally relevant information in the language and modality people are familiar with, such as radio or simple voice-based phones.

In the Tiballi project, rainfall data are collected in the field and combined with global weather and climate information available on the Web. The combined result is made available to the local community members in their language, through voice-based phone access or local radio (Baart et al., 2018). Automatic voice response and interaction is constructed from vocabularies which are collected in the community and processed with AI. Figure 2 shows a focus group discussion with farmers in the Tingoli village in northern Ghana.

In short, to deliver the requested information to the community, the Tiballi project aims to:

(i) Collect and store real-time weather, for example, the cumulative rainfall over the season.
(ii) Develop an automatic speech recognition system in the local language Dagbanli. It uses a previously developed method to resource small languages,

in which a machine learning model is trained with a relatively small, locally collected dataset.

(iii) Develop a voice-based application that combines limited vocabulary text-to-speech and automatic speech recognition in Dagbanli language. This will enable users to interact with the system by phone and receive requested information, every day (Stan et al., 2022).

(iv) Develop and deploy a voice-based platform, based on inexpensive, small hardware, to provide the service, also in the absence of Internet access and avoid high costs of broadband connectivity (Baart et al., 2019).

7 Discussion on Critical and Societal Issues

In the previous sections, we proposed a collaborative, iterative, adaptive approach to digital development, aligned with inclusion and an ethics perspective. Yet, this approach may have its limitations, for example, related to sustainability or costs of digital services. It is important to continuously evaluate and critically reflect before, during, and after the project period. Deliberation and discussions with its users and beneficiaries on their personal experiences and opinions are essential.

Technological innovation can be a rapid process. Huge developments are taking place in (distributed, cloud, edge) computing, social media, blockchain technology, Internet of Things, artificial intelligence, and data science. It is important to continuously reassess potential positives and negatives of technologies introduced in low-resource environments. We as researchers, ICT professionals, and global citizens must be aware of our responsibility, to care for the present and the future of the digital society.

From the ideas in this chapter, we propose to support the lives and work of people in low-resource environments through co-design. Yet, this small-scale approach needs dissemination to scale up, for example, by training young professionals at European and African universities and by bringing together a community of developers and contributors.

For a societally oriented community of developers, it may be challenging to deliver, in a sustainable way, competitive digital services to low-resource environments. The strong monopolization by big technology firms with technologically advanced services, and backed by big investors, puts up high barriers for the smaller parties. Yet, niches exist—especially in less wealthy regions—that are still unserved.

This endorses the idea that deliberate choices can be made toward a better and more inclusive digital society, guided by social engagement, and encourage human-centered innovation, quoting the Vienna Manifesto 2019: “calling for a Digital Humanism that describes, analyzes, and, most importantly, influences the complex

interplay of technology and humankind, for a better society and life, fully respecting universal human rights"[4] (DIGHUM, 2015).

8 Conclusions

Whereas, from the previous discussions, it is evident that, from a macro-economic and geopolitical perspective, there are no easy fixes to bridge the digital divide as a problem of global scale between regions and countries, we propose to develop small-scale grassroots approaches. We stress the importance of collaboration, at all levels.

Firstly, at the international policy level—in platforms and think tanks such as the Digital Humanism Initiative—representatives of communities in the least connected countries and regions in the Global South must be included, to make their perspectives visible, create awareness of the existing problems, and inform the public debate about the needs of people in low-resource environments.

Secondly, at the level of research and education, collaboration and exchange of knowledge is needed between universities (north-south and south-south). Research programs must be developed that are context oriented. Students must be trained through community service-oriented research and technology development.

Finally, it is important to innovate in technologies in and for people in low-resource environments and develop transdisciplinary research methodologies. In the paragraphs above, we discussed community-oriented, collaborative action research and socio-technical software development. While the exchange of ideas between developers and users can offshoot innovation in low-resource environments in unexpected ways, it can also be a source of inspiration for new forms of collaborative knowledge production.

Discussion Questions for Students and Their Teachers

1. What does it mean when it is said that the digital space is *colonized*?
2. What are the problems when digital technologies are transferred from industrialized countries to developing countries, and what would be the alternatives to digital development?
3. What are the decolonial aspects of digital technologies that reflect patterns from the social and physical world? Think of examples that are not mentioned in this chapter.
4. What is the fundamental difference between the critical realism perspective and the action research/design science perspective on the challenges related to the digital divide?
5. What are the five main contextual challenges for the design of digital services, of the low-resource environment in the Tiballi case study in northern Ghana, and how does the given framework ICT4D 3.0 address these challenges?

[4]This quote is from the Vienna Manifesto on Digital Humanism, May 2019.

Learning Resources for Students

1. Potter, R. B., Binns, T., and Elliott, J. A. (2008) *Geographies of development: An introduction to development studies*. Essex: Pearson Education. Chapter 1.

   This book gives, from a geopolitical perspective, an introduction to the digital divide between countries and regions in the world and describes the historical backgrounds in the framework of recent and contemporary history.
2. Zuboff, S. (2019) The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power. London: Profile books.

   This book provides a critical analysis of the business models used by big technology firms and how these affect people and the economy, creating unfairness in the digital society and the physical world.
3. Mohamed, S., Png, M. T. and Isaac, W. (2020) Decolonial AI: Decolonial Theory as Sociotechnical Foresight in Artificial Intelligence. *Philosophy & Technology,* 33, pp. 659–684.

   This paper analyzes the risks and biases of artificial intelligence algorithms through the lens of decolonial theory.
4. Stan, G. et al. (2022) 'A Lightweight Downscaled Approach to Automatic Speech Recognition for Small Indigenous Languages', in *14th ACM Web Science Conference 2022* (pp. 451–458). Doi: https://doi.org/10.1145/3501247.3539017.

   This paper describes a technical method to build AI-based automatic speech recognition for indigenous languages, using small data and energy-efficient methods, making it more affordable and accessible for people and communities in low-resource environments,
5. Dittoh, F., et al. (2021). 'Tibaŋsim: Information Access for Low-Resource Environments', in: *Conference proceedings by Springer AISC*. ISBN Number—2194-5357 Series. Doi: https://doi.org/10.1007/97898116.2377662**.**

   This paper describes a real-world deployment of an ICT system in rural Ghana, as an example of collaborative technology design and its challenges in low-resource environments in the Global South.
6. Baart, A., et al. (2018) 'Ney Yibeogo—Hello World: A voice service development platform to bridge the web's digital divide', in *WEBIST 2018—Proceedings of the 14th International Conference on Web Information Systems and Technologies,* pp. 23–34.

   This paper stresses the importance of resourcing local languages to bridge the digital divide and methods how to do this.

Acknowledgment The case study of the Tiballi research project, described in this section, is funded by the Internet Society through a research grant in 2023. Also, the authors have received some funding from the EU program Erasmus+ for the field visits and the elaboration of educational modules, which have resulted in this chapter.

References

Augustine, A. (2022). *The next wave: A data centre roadmap for Africa.* Accessed June 22, 2023, from https://techcabal.com/2022/10/03/data-centre-africa/.

Baart, A., et al. (2018). Ney Yibeogo—Hello World: A voice service development platform to bridge the web's digital divide. In *WEBIST 2018—Proceedings of the 14th international conference on web information systems and technologies* (pp. 23–34).

Baart, A., et al. (2019). Affordable voice services to bridge the digital divide: Presenting the Kasadaka platform. In *Selected papers from the WEBIST 2018 international conference*. Springer LNBIP Book Series.

Berry, J. (2006). The world summit on the information society (WSIS): A global challenge in the New Millennium. *Libri, 56*(1), 1–15. https://doi.org/10.1515/LIBR.2006.1

Bon, A. (2016). ICT4D 3.0: An adaptive, user-centred approach to innovation for development, In *Conference on advanced information systems engineering*, CAiSE 2016.

Bon, A. (2020). *Intervention or collaboration? Redesigning information and communication technologies for development*. Pangea. https://doi.org/10.26481/dis.20201215ab

Bon, A., et al. (2022). Decolonizing technology and society: A perspective from the global south. In H. Werthner et al. (Eds.), *Perspectives on digital humanism*. Springer. https://doi.org/10.1007/978-3-030-86144-5_9

Buchele, S. F., & Owusu-Aning, R. (2007). The one laptop per child (OLPC) project and its applicability to Ghana. In *Proceedings of the 2007 international conference on adaptive science and technology* (pp. 113-118).

Dahir, A. L. (2020). A bird? A plane? No, It's a Google Balloon beaming the Internet. Accessed June 22, 2023, from https://www.nytimes.com/2020/07/07/world/africa/google-loon-balloon-kenya.html.

De Janvry, A., & Sadoulet, E. (2021). *Development economics: Theory and practice* (2nd ed.). Routledge.

De, R., et al. (2018). ICT4D research: A call for a strong critical approach. *Information Technology for Development, 24*(1), 63–94.

DIGHUM. (2019). *Vienna Manifesto on digital humanism*. Accessed June 23, 2023, from https://dighum.ec.tuwien.ac.at/dighum-manifesto/.

Dittoh, F., et al. (2020a). Mr. Meteo: Providing climate information for the unconnected. In: *Companion publication of the 12th ACM conference on web science*. (pp. 20-25). https://doi.org/10.1145/3394332.3402824.

Dittoh, F., et al. (2021). Tibaŋsim: Information access for low-resource environments. In *Conference proceedings by Springer AISC*. ISBN number—2194-5357 series. https://doi.org/10.1007/97898116.2377662.

Dittoh, F. et al. (2020b). Information access for low-resource environments. In *Proceedings of the 3rd ACM SIGCAS conference on computing and sustainable societies*. (pp. 325-326).

Fuchs, C., & Horak, E. (2008). Africa and the digital divide. *Telematics and Informatics, 25*(2), 99–116.

Grewal, D. S. (2008). *Network power: The social dynamics of globalization*. Yale University Press.

Heeks, R. (2008). ICT4D 2.0: The next phase of applying ICT for international development. *Computer, 41*(6), 26–33.

Hope, J. (2022). Globalising sustainable development: Decolonial disruptions and environmental justice in Bolivia. *Area, 54*(2), 176–184.

Kalba, K. (2008). The adoption of mobile phones in emerging markets: Global diffusion and the rural challenge. *International Journal of Communication, 2*, 31.

Lin, C. I. C., Kuo, F. Y., & Myers, M. D. (2015). Extending ICT4D studies: The value of critical research. *MIS Quarterly, 39*(3), 697–712.

Mbiti, I., & Weil, D. (2011). *Mobile banking: The impact of M-PESA in Kenya*. US National Bureau of Economic Research. NBER Working Paper No. 17129.

Mendoza, B. (2021). Decolonial theories in comparison. In S. M. Shih & L. C. Tsai (Eds.), *Indigenous knowledge in Taiwan and beyond. Sinophone and Taiwan studies* (Vol. 1, pp. 249–271). Springer. https://doi.org/10.1007/978-981-15-4178-0_12

Mignolo, W. D., & Walsh, C. E. (2018). *On decoloniality: Concepts, analytics, praxis*. Duke University Press.

Mohamed, S., Png, M. T., & Isaac, W. (2020). Decolonial AI: Decolonial theory as sociotechnical foresight in artificial intelligence. *Philosophy & Technology, 33*, 659–684.

Mubarak, F. (2015). Towards a renewed understanding of the complex nerves of the digital divide. *Journal of Social Inclusion, 6*(1), 71–102.

Nan, W. V. (2019). Mobile money and socioeconomic development: A cross-country investigation in Sub-Saharan Africa. *Journal of International Technology and Information Management, 27*(4), 36–65.

Pini, M. (2020). Digital inequality in education in Argentina: How the pandemic of 2020 increased existing tensions. In *12th ACM conference on web science companion* (pp. 37-40).

Potter, R. B., Binns, T., & Elliott, J. A. (2008). *Geographies of development: An introduction to development studies*. Pearson Education.

Prahalad, C. K. (2004). *The fortune at the bottom of the pyramid: Eradicating poverty with profits*. Wharton Business Publishing.

Quijano, A. (2007). Coloniality and modernity/rationality. *Cultural Studies, 21*(2–3), 168–178.

Qureshi, S. (2006). Why is the information society important to us? *The World Summit on the Information Society in Tunis, 12*(1), 1–5. https://doi.org/10.1002/itdj.20035

Rogers, E. M. (2003). *Diffusion of innovations* (5th ed.). Free Press.

Rolland, C. (2007). Capturing system intentionality with maps, In *Conceptual modelling in information systems engineering,* Springer, (pp. 141-158).

Sanchez, P., et al. (2007). The African millennium villages. *Proceedings of the National Academy of Sciences, 104*(43), 16775–16780.

Sapah, M. S. (2023). *Starlink: SpaceX's new internet service could be a gamechanger in Africa.* Accessed June 22, 2023, from https://theconversation.com/starlink-spacexs-new-internet-service-could-be-a-gamechanger-in-africa-200746.

Stan, G. et al. (2022). A lightweight downscaled approach to automatic speech recognition for small indigenous languages. In *14th ACM web science conference 2022* (pp. 451-458). https://doi.org/10.1145/3501247.3539017.

Taylor, J. E., & Lybbert, T. J. (2020). *Essentials of development economics*. University of California Press.

The Guardian. (2016). Facebook lures Africa with free internet—but what is the hidden cost? Accessed June 22, 2023, from https://www.theguardian.com/world/2016/aug/01/facebook-free-basics-internet-africa-mark-zuckerberg.

Tiballi. (2023). *Making AI work for Inclusiveness in the Global South.* Accessed June 22, 2023, from https://tiballi.net/.

Van Dijk, J. (2020). *The digital divide*. Polity Press.

W4RA. (2023) *The Web alliance for regreening in Africa.* Accessed June 22, 2023, from https://w4ra.org/.

Wicander, G. (2009). The invisible hand of the market or the visible hand of the state—how to reach universal access? In J. S. Pettersson (Ed.), *Defining the 'D' in ICT4D: Graduate papers on development, Globalisation and ICT*. Karlstad University.

Zuboff, S. (2019). *The age of surveillance capitalism: The fight for a human future at the new frontier of power*. Profile Books.

SPEAKER

Second African Digital Humanism Summer School, Accra, Ghana, 31 August – 05 September 2026

Peter Knees

Peter Knees is a Full Professor of the Faculty of Informatics, TU Wien, Austria and a Visiting Assistant Professor at Georgia Institute of Technology, School of Music for the Fall term 2022. He holds a Master's degree in Computer Science from TU Wien and a PhD in the same field from Johannes Kepler University Linz, Austria. For almost two decades, he has been an active member of the Music Information Retrieval research community, reaching out to the related fields of multimedia and text information retrieval, recommender systems, and the digital arts. His research activities center on music search engines and interfaces as well as music recommender systems, and more recently, on smart(er) tools for music creation. He is one of the proponents of the Digital Humanism initiative of the Faculty of Informatics of TU Wien. Since 2023, he has also headed the UNESCO Chair in Digital Humanism.

Lecture 4: *Artificial Intelligence - & Where It Will/Might/Could/Should Go*

Reader article:

Peter Knees: Scaling Up Broken Systems? Considerations from the Area of Music Streaming. In: Werthner, H., Prem, E., Lee, E.A., Ghezzi, C. (Eds.): Perspectives on Digital Humanism, pp 165-171. Springer Nature, Cham, Switzerland, 2022. https://doi.org/10.1007/978-3-030-86144-5_23

Scaling Up Broken Systems? Considerations from the Area of Music Streaming

Peter Knees

Abstract We discuss the effects and characteristics of disruptive business models driven by technology, exemplified by the developments in music distribution and consumption over the last 20 years. Starting from a historical perspective, we offer insights into the current situation in music streaming, where technology has not only changed the way we access music but also has important implications on the broader ecosystem, which includes the consumers, the authors, the record industry, and the platforms themselves. The discussion points to potential benefits, as well as to the risks involved in the currently deployed systems. We conclude that the increased profitability of the disruptive business models in the music domain and beyond is largely generated at the expense of the providers of the goods or services being brokered. Using the platforms as a consumer further subsidizes their value and might lead to mono- and oligopolies. While technology allows companies to effectively scale up business, the resulting systems more often amplify existing injustices than mitigate them.

1 Introduction

Technological developments have always created opportunities for business. This is especially the case for disruptive technologies. In the past decades, we have increasingly seen a fascination of venture capital with businesses promising disruption in their area and therefore maximizing investment return. Some popular examples of this paradigm include platforms for retail (Amazon, Alibaba), accommodation (Airbnb, Booking.com), individual transportation (Uber, Lyft), and media distribution (Netflix, Spotify) or, e.g., fintech services (Klarna, TransferWise). Characteristic to these business models across all domains is some form of incentive, such as improved user experience and/or reduced cost, for the various stakeholders involved, and increased efficiency by automating central elements of the traditional business

P. Knees (✉)
Vienna University of Technology, Vienna, Austria
e-mail: peter.knees@tuwien.ac.at

H. Werthner et al. (eds.), *Perspectives on Digital Humanism*,
https://doi.org/10.1007/978-3-030-86144-5_23

approach, therefore allowing to scale up easily. These advantages undoubtedly are a key element for the success of the new business forms and overall the Internet-driven economy. Looking at the old-fashioned, obviously imperfect, and ultimately replaced business models reveals many flaws and inherent injustices. However, comparing them with their disruptive successors, it shows that the stakeholders consisting of a group of individual, non-organized participants (the providers of goods, as well as the consumers) tend to further lose power and influence in favor of the business owners providing the service. This raises the question: are these digital technologies merely permitting to scale up systems that were broken from the beginning? And if so: are disruptive digital technologies just amplifying injustices of the prior status quo?

In the following, we ask this question from the perspective of the area of music distribution. Music played an early and prominent role in the digital transformation of media delivery and consumption and has undergone significant changes over the last 25 years from a brick-and-mortar retail good to a ubiquitously available, de facto free service. We aim at giving a historical perspective describing the situation and the different stakeholders involved prior and past the disruption of the business and give an account on the beneficiaries and benefactors of this process, before drawing parallels to other areas of business disruption.

2 A Brief History of Music Business

Historically, the development of music production, distribution, and consumption is tightly intertwined with the technical possibilities of their respective time. Consider, for instance, the consequences of being able to transmit, physically record, and play back acoustic signals and the resulting means of broadcasting music and decoupling performance and listening. Over the second half of the twentieth century, this has led to the establishment of a record industry that would shift focus and power away from the composers of music and the publishing of sheet music toward the performers and marketing of events and trends. Revenue in this system was mostly generated by selling physical records on various types of media, most prominent on vinyl, music cassette, and compact disc (CD), with various stakeholders (among others, composers, performers, record labels, publishers, producers, distributors, and retailers) ideally being remunerated according to their negotiated share. Other, comparatively minor sources of revenue would stem from royalty payments based on airplay, performance, and licensing, as well as concerts. Radio was (and is) primarily a marketing tool to promote artists and records. To ensure payment to the various stakeholders, royalty collecting agencies established themselves to act on behalf of the rights owners, setting up rules for money distribution based on measurable, albeit ultimately arbitrary criteria, such as chart positions and other industry-conceived mechanisms. Essentially, this yielded a division of the area of music into "superstars" that could easily claim a (large, actually overproportioned) share and those who would have to fight for and often yield theirs.

Overall, this system favored the record companies over the individual artists and creators, as controlling the music production means would allow them to exert power over artistic decisions and aggregate various licensing rights while maintaining strong ties to the media and other distribution channels. In practice, the market was controlled by the major record labels and a number of gatekeepers, establishing an oligopoly that works for a few at the core of the industry at the expense of the many contributors. As a side effect of this model, however, record companies would also afford to finance the development of new artists, i.e., to invest in artists and repertoire (A&R). For independent artists, the diversity of media outlets and distribution channels would nonetheless allow them to find niches and address potential audiences.

3 The Disruption of the Business

The end of the twentieth century brought new technological developments, again impacting the music industry—and again, before all others. The adoption of the Internet as communication medium and the invention of audio data compression algorithms and formats, specifically MP3, led to active exchange of digital music files among users via decentralized peer-to-peer networks such as Napster, outside all established channels, infrastructures, and monitoring mechanisms built by the music industry. Needless to say, the resulting, almost instant availability of comprehensive music repositories to every peer of the network posed a threat to the core business model of the industry. Instead of embracing the new technological opportunities, the developers and users of the networks were criminalized and the downfall of popular musicians and music as an artform itself foretold as the music industry would not be able to keep up the monetization strategy that supported all stakeholders, related businesses, and intermediaries involved. As a reaction, the entertainment industry successfully lobbied for changes in copyright and intellectual property rights laws to further strengthen their position.

A few years into these developments, which indeed brought declining profits for the music ecosystem in comparison to the peak of CD sales in 2001 (IFPI 2021), new players and services explored the potential of the informatization of music reproduction and the decoupling from a physical media in alternative business models. These ranged from traditional e-commerce models of digital stores for music files to newer paradigms such as flat-rate subscriptions or freemium models typically with an ad-based entry level for unlimited music streaming. Together with the rise of smartphones that pushed an “always online” philosophy, new forms of access to music were enabled.

Generally, this framework of digital distribution via online services offers many advantages over the historic approach and potential for mitigations to its shortcomings for the various stakeholders. For music consumers, instant, cheap, and legal access to large amounts of music becomes feasible. For music creators, including those who contribute to the production process, it can instantly provide traceable

delivery of their music to their audiences without having to deal with the various gatekeepers and distribution hubs. For the industry in between, logistics of physical retail is no longer necessary. For the music platform itself, being the central hub for delivering music to users allows them to track music listening behavior and build personalized services such as recommender systems and advanced interfaces for music discovery (Knees et al. 2020). In this scenario, however, record labels might lose their position of power and be bypassed easily. Their bargaining chip to retain relevance in this development is the control over the back catalogues, i.e., licensing access to records by well-known and popular artists whose availability is imperative for broad adoption. Ultimately, the disruption of the business introduced streaming platforms as additional central market players, next to record companies, whereas brick-and-mortar retail has largely vanished.

4 The Status Quo

While sales of physical media have been declining since 2001, the rising revenue from streaming has almost compensated for the losses and delivered an overall turnaround in 2015, today constituting the largest source of revenue (IFPI 2021). With more than 440 million paid subscription users worldwide (ibid.), streaming is the dominating modality of (traceable) music consumption. Current catalogues of the market-leading streaming platforms such as Spotify, SiriusXM Pandora, Apple Music, Amazon Music, or YouTube Music offer up to 70 million music tracks (Spotify 2021). To sum up, music streaming has managed to scale up the music business both in terms of content accessible and users reached.

But how is this situation still (or even more?) profitable with more players to support? The payments made by streaming companies for individual streams are extremely small—on average about USD 0.004 per stream (see Pastukhov (2019))—and often even negligible in sum per artist. Moreover, this amount is paid to the owners of the master recordings, which are typically the record labels, which then pass on only a small fraction to the artists. A reason for this practice is that revenue from streaming is often not explicitly negotiated in contracts, particularly in those signed before the technology existed. Hence, record companies benefit from licensing their back catalogues as it allows them to keep the largest share of the royalty payments to themselves. Also, recent efforts to present more justified and "fair" money distributions models, e.g., by distributing subscribers' fees according to their individual listening preferences, as implemented in Deezer's user-centric payment system (Deezer 2021), are not improving the situation as again the money is not paid directly to the artists, but the producers and publishers. Furthermore, despite the technical possibility to trace all plays of individual tracks, information about play counts across various platforms remains opaque for music creators, especially on platforms driven by user-generated content like YouTube on which content identification focuses again on the needs of big rights owners and neglects independent artists. These shortcomings themselves led again to the establishment of disruptive

businesses like Kobalt or UnitedMasters that combine the roles of independent labels, rights management, and publishing companies, providing easier access to platforms and increased transparency of the process in a unified interface.

Record companies can now focus their attention and marketing strategies to a few distribution channels instead of dealing with a variety of different outlets. Artists and repertoire (A&R) is not part of their core business anymore, as new trends and talents are observed on social media platforms and co-opted based on hypes. As streaming is not generating much revenue, alternative sources of income like live performances, merchandise, sponsoring, and licensing of tracks need to provide a stable financial foundation for artists. At the same time, record labels increasingly demand a share of this profit in exchange for managing and promoting artists.

While record labels have once more found their position at the center of the business at the expense of artists and music creators, the question of profitability of the streaming platforms remains. Based on the annual reports of the various platforms, it appears they are not (yet?) operating profitable businesses. Their value seems to lie elsewhere, namely, in the potential for commercially exploiting their user basis, not only by charging for the service, but in fact by providing a data basis for marketing analyses sold to brands and advertising (Zuboff 2018). This might explain why, for instance, Spotify is partly owned by an intricate structure of the biggest record companies and rights holders, consisting of Universal Music Group, Warner Music Group, Sony Music, and, most recent, Tencent (Ingham 2020; see also Eriksson et al. 2019).

5 . . . And Beyond?

While in the music business the old industry reinvented their role and managed to ultimately shape the new music industry based on its assets and absorb the disruptive element, in other domains, disruption has led to a replacement of the intermediaries and gatekeepers. In the movie domain, Netflix, first as a recommender-driven mail-in service, then as a movie streaming pioneer, completely eliminated the established video and DVD rental business and opened up a market that is now a battleground for several competitors following the same model. In the transportation area, Uber entered a long-time undisputed market and took large shares by storm. In the tourism domain, platforms like Booking and Expedia pushed the traditional business model of travel agencies close to the edge of the cliff (while the COVID-19 crisis seemingly finally pushed it over).

While Web-based services are at the center of all these examples and provide a new dimension of user experience in all these domains, it is never the user side that makes the business profitable. In music, the changed industry landscape is carried by besting the composers and creators of music due to old contracts or reduced benefits. In transportation, the pressure is put on drivers to make up for reduced wages by increasing customer throughput. In the tourism domain, hotels are put under further pressure to reduce their own profit margins and undersell their capacities to get

visibility on platforms. In the movie domain, on the other hand, the competition of streaming services and their newly founded studios seems to increase the opportunities within the movie and TV industry rather than forcing the producers and workers to produce more for less compensation. In this context, it should be noted that the movie industry, especially in the USA, is strongly unionized.

The user is not spared from possible negative consequences either. Resorting again to the music example, the status quo of having access to virtually all music all the time is a utopian situation for music aficionados, and the technical developments that have enabled this situation over the last decades are a shining example of the blessings of information technology. At the same time, users become strongly dependent on and locked into such platforms. On one hand, "having access" does not mean "to own," and "personal collections" consisting of playlists created on platforms might become incomplete or missing due to changed circumstances in licensing. On the other hand, such services become more indispensable the more they are used. This might lead to the more general observation that "the rise in market concentration is greater in industries that are more intensive users of digital technologies" (Qureshi 2019); i.e., the disruption of industries by means of digital technologies itself promotes the emergence of mono- or oligopolies. In the long run, this is disadvantageous for the user and most other stakeholders involved.

To conclude, technology drives and "sells" disruptive business models. The increased profit of these models is often and largely generated by exploiting the people who provide the actual value behind the product, especially if they are not organized to fight for their joint interests. The users further subsidize these businesses simply through usage. From the perspective of Digital Humanism, this situation is unsatisfactory. We should first use technology to overcome the many broken systems before helping scaling them up.

References

Deezer (2021). Deezer wants artists to be paid fairly. Online: https://www.deezer.com/ucps, retrieved: 05/02/2021.

Eriksson, M., Fleischer, R., Johansson, A., Snickars, P., & Vonderau, P. (2019). Spotify Teardown: Inside the Black Box of Streaming Music. Cambridge: MIT Press.

IFPI (2021). Global Music Report 2021. Online: https://gmr2021.ifpi.org/assets/GMR2021_State%20of%20the%20Industry.pdf, retrieved: 05/02/2021.

Ingham, T. (2020). Who Really Owns Spotify?. Rolling Stone. Online: https://www.rollingstone.com/pro/news/who-really-owns-spotify-955388/, retrieved: 05/02/2021.

Knees, P., Schedl, M., & Goto, M. (2020). Intelligent User Interfaces for Music Discovery. Transactions of the International Society for Music Information Retrieval, 3(1), pp.165–179. DOI: https://doi.org/10.5334/tismir.60

Pastukhov, D. (2019). What Music Streaming Services Pay Per Stream (And Why It Actually Doesn't Matter). Soundcharts. Online: https://soundcharts.com/blog/music-streaming-rates-payouts, retrieved: 05/02/2021.
Qureshi, Z. (2019). Inequality in the Digital Era. Work in the Age of Data. BBVA OpenMind Collection (12), BBVA.
Spotify (2021). Company Info. Online: https://newsroom.spotify.com/company-info/, retrieved: 05/02/2021.
Zuboff, S. (2018). The Age of Surveillance Capitalism: The Fight for the Future at the New Frontier of Power. London: Profile Books.

Digital Humanism Initiative
Technology for Humanity

SPEAKER

Second African Digital Humanism Summer School, Accra, Ghana, 31 August – 05 September 2026

Maud Adjeley Ashong Elliot

Maud Adjeley Ashong Elliot is a computer scientist extraordinaire who encourages female students to enter the field of Information Technology. At the Department of Information Technology Studies at the University of Professional Studies, Accra, she teaches Systems Analysis and Design and Management Information Systems. She is an IBM-trained educator in mobile and cloud development. Graduated from the Kwame Nkrumah University of Science and Technology, she started her teaching career at the Computer Science Department of the same university. Her research interests are digital transformation and women in technology. She is an advocate for safe internet usage for young & old and she advocates and supports for early teaching of programming languages such as Python to children.

Lecture 5: *Digital Technologies and Policies in the African Public Sector*

Reader article:

Maud Adjeley Ashong Elliot, John Effah, and Richard Boateng: Institutional Influence on e-Justice System Usage: The Case of Ghana's Judicial Sector.
In AMCIS 2024 (Thirtieth Americas Conference on Information Systems, Salt Lake City, 2024) Proceedings, 28. Association for Information Systems AIS, Atlanta, GA, USA, 2024.
https://aisel.aisnet.org/amcis2024/adoptdiff/adoptdiff/28

August 2024

Institutional Influence on e-Justice System Usage: The Case of Ghana's Judicial Sector

Maud Adjeley Ashong Elliot
University of Professional Studies, Accra, maud.ashongelliot@upsamail.edu.gh

John Effah
University of Ghana, jeffah@ug.edu.gh

Richard Boateng
University of Ghana, richboateng@ug.edu.gh

Institutional Influence on e-Justice System Usage: The Case of Ghana's Judicial Sector

Completed Research Full Paper

Maud Adjeley Ashong Elliot
University of Professional Studies, Accra
maud.ashongelliot@upsamail.edu.gh

John Effah
University of Ghana Business School
jeffah@ug.edu.gh

Richard Boateng
University of Ghana Business School
richboateng@ug.edu.gh

Abstract

This study seeks to understand how institutional forces —cultural-cognitive, regulative, normative— influence e-Justice system usage in the judicial sector of Ghana. Utilizing an interpretive case study approach and institutional theory as a lens, it explores how these institutional pillars shape the adoption and effectiveness of the e-Justice system within Ghana's judicial sector. The findings reveal that regulatory directives from judicial leadership, normative pressures reflecting global digitalization trends, and cognitive factors related to stakeholders' perceptions and willingness to adopt technology significantly impact the system's utilization. This research contributes to the understanding of digitalization in developing countries' judicial systems, highlighting the complex interplay between technology and institutional forces.

Keywords

Institutional forces, digitalization, e-Justice System, interpretive case study, Ghana.

Introduction

In recent years, the global landscape has shifted from a traditional setting to a digital one. Information and communication technologies keep changing how we live and interact in society (Salah-Eddine 2022). As a result, adopting digitalization initiatives has become a significant phenomenon across many countries and sectors (Kuusisto 2017). Information system literature reveals that digitalization efforts have increased due to the COVID-19 pandemic (Danyliuk et al. 2021; Martínez et al. 2023; Tungpantong et al. 2021). Digitalization refers to the "sociotechnical process of applying digitizing techniques to broader social and institutional contexts" (Tilson et al. 2012). Digitizing techniques are the technical processes of converting data from analogue into digital forms (Tilson et al. 2010). Thus, the phenomenon encompasses more than just the conversion of analogue information to digital format but also emphasizes the transformation of organizations' processes to provide new value propositions (Bloomberg 2018; Reis et al. 2020). Digitalization is a significant organisational transformation activity undertaken to enhance efficiency, flexibility, and creativity (Parida et al. 2019; Yoo 2010). Digitalization initiatives have been seen in sectors like education (Bygstad et al. 2022), finance (Tsindeliani et al. 2022), healthcare (Preko et al. 2019), manufacturing (Liere-Netheler et al. 2018), and pharmaceuticals (Okano et al. 2021). In the digital age, incorporating technology into judicial processes is crucial for improving legal proceedings, increasing access to justice, and assuring efficient case management (Jneid et al. 2019). Ghana, a West African country dedicated to improving its public sector, is making strides in the Digitalization of its judicial sector (Bawole and Langnel 2022).

The digitization of the judicial system aims to enhance the transparency and accessibility of justice for every citizen (Đilas and Kuzmanović 2023). In the realm of the judiciary, digitization initiatives might include leveraging technology to improve numerous aspects of the legal process, including but not limited to, the electronic submission of documents, digital payment of legal fees, virtual dispute resolution, digital case management, remote hearings, and trials, as well as legal research among others (Loo and Findlay 2022;

Shevchenko 2022). This research is particularly valuable for stakeholders within Ghana's Judicial sector (i.e administrative staff, judiciary, Ministry of Justice and all who are responsible for maintaining rule of law) aiming to enhance judicial services through technology, offering a detailed examination of the dynamics at play which could guide the design and implementation of similar systems in other developing countries. Although the integration of digitalization within the public sector is on the rise (Drossos et al. 2018), there has been limited research concerning the phenomenon within the judicial sector of developing or emerging countries (Hanelt, Bohnsack, Marz, & Antunes 2021). Again, the few existing studies on digitalization in the judicial sector from developing countries including Egypt (Aboelazm 2019), Morocco (Salah-Eddine 2022), and Cape Verde (Rosa et al. 2013) as atheoretical studies have focused on the adoption, implementation, success factors, challenges, and impact of the e-Justice system. There is a need therefore for studies that employ theories that can discuss the influence of institutional context on the usage of the systems. In line with this empirical gap, this study therefore also seeks to extend the literature on the digitalization of the judicial sector with evidence from Ghana. Ghana is used as the context for this study because it is a developing country and has attempted to electronically transform its public services after the COVID-19 pandemic. By adopting the interpretive case study approach (Klein and Myers 2001; Walsham 1995) and using institutional theory (DiMaggio and Powell 1983; Scott 2004) as the theoretical foundation, this study aims to comprehend the influence of institutional pillars on the use of the e-Justice system in the judicial sector of Ghana. It is worth noting that, to the best of the knowledge of the researchers, this is the first study on e—Justice systems adopting institutional theory as a lens in a developing country context. The research question motivating this study therefore is, *how do the regulative, normative, and cultural cognitive institutional pillars shape e-Justice system usage in Ghana's judicial sector?*

This study will therefore explain how regulative, normative, and cultural-cognitive institutional forces influence e-Justice system usage in Ghana's judicial sector, providing a novel application of institutional theory on e-Justice systems studies in a developing country context. It offers practical insights and policy recommendations on managing digitalization in judicial systems, informed by a detailed analysis of global digitalization trends and their local impacts.

The organization of the manuscript follows a sequential format as outlined: literature on digitalization of the judicial sector and e-Justice system, the theoretical foundation where Institutional theory is adopted as the lens, the research methodology, the findings, analysis, and discussions, and finally the conclusion.

Digitalization of Judicial Sector through e-Justice System

Interest in digitalization has been growing among IS researchers (Bharadwaj et al. 2013) and practitioners (Fitzgerald et al. 2013) alike. Operating independently within the government, the judicial sector's primary functions include the administration and upholding of law and order and the protection of human rights. It engages in a judicial process that necessitates various types of data collection and interaction among actors. This process, encompassing the filing and the exchange of documents and paperwork, often leads to significant and unnecessary prolongation in justice provision due to its conventional practices and procedures (Murphy and Pritchett 1986). Therefore, the judicial sector must digitize its processes, ensuring the involvement of all relevant stakeholders.

Electronic justice (e-Justice) involves applying digital technologies to enhance the judicial process within a specific domain (Sandoval-Almazan and Gil-Garcia 2020). Several authors endeavored to describe e-Justice systems on a jurisdiction basis. For instance, Estonia has been distinguished as a front-runner in transforming the Judicial sector. The e-Justice system there has improved legal procedures by providing a centralized platform for information access, document submission, and case tracking. Also, emerging technologies like AI and blockchain are incorporated into the system to aid in speech-to-text transcriptions, the use of "robotic judges" in adjudicating small claims, and enhancing the security of electronic documents. However, the authors indicate that there is a need for legislation to regulate the usage of these technologies in courtrooms, especially with the usage of robotic judges (Marta and Giulia 2023). Furthermore, Ivanovna's (2020) study on e-Justice in Russia found that the President decreed the task in 2018, and the country has implemented systems in the Supreme Court and Court of Arbitration dubbed "My Arbitr and Provisudae", accepting both non-electronic and electronic (i.e. pdf, jpg., docx) documents in legal proceedings. Moreover, a study in Ukraine revealed that the e-Justice System, the Unified Judiciary Information and Telecommunication System (UJITS), made it less stressful for citizens to access court information and services. UJITS enabled the utilization of video conferencing for virtual hearings,

electronic filing and submission of documents, and remote signature. However, the challenge indicated was the fact that some Judges, Lawyers, and other stakeholders were initially reluctant to use the system because they were not tech-savvy. Aboelazm (2019) recounts the experience of Egypt. He disclosed that e-Justice has contributed to achieving prompt justice in Egypt in the Economic Court through the integrated modules including e-filing, e-case progress, and e-record among others. However, he suggests that e-Justice should be scaled up to include e-payment of fees modules and be deployed to the courts of the state council at all levels. Moreover, an investigation into Cape Verde's e-Justice system reveals that ensuring the successful usage of the e-Justice system is dependent upon the quality of infrastructure and enabling stable internet connectivity precisely in remote areas(Rosa et al. 2013). Finally, a study on the digitalization of the judiciary in Iran by (Kondori and Rouhani 2021) found that realizing the full potential of technology hinges on the interplay of culture, skills, and legal frameworks. These findings underscore the reality that digitalization brings forth several challenges. Some dominant challenges pinpointed in the literature were funding, legislation, cultural factors, training requirements, staff capacity, and matters related to data privacy, integrity, and accessibility (Jneid et al. 2019; Rosa et al. 2013). Again, these findings show that culture, law, and institutional environment could influence the potential of e-Justice systems. This study, therefore, extends the dialogue on e-Justice systems to analyze the institutional pressures that shape or fluence usage from the Ghanaian context.

Theoretical Foundation

This study draws on institutional theory (DiMaggio and Powell 1983) as a theoretical foundation to understand the "regulative, normative, and cognitive" institutional influence on e-Justice usage in Ghana. Institutional theory is regarded as a valuable framework for understanding how existing social frameworks, such as norms, regulations, and cognitive processes, influence social behaviors and interactions (Aksom and Tymchenko 2020; Dacin et al. 2002). Thus, institutions represent entrenched legal and socio-cultural norms that direct the behavior of individuals and organizations within a specific social setting (Avgerou 2000; DiMaggio and Powell 1983). According to (Scott 2004) there are three pillars underpinning institutional theory; these include, regulative, cognitive, and normative.

Initially, regulative institutions are defined as government mandates and industry benchmarks that organizations must adhere to; non-compliance results in penalties (Kshetri 2007). Regulative institutions thus dictate what actions are deemed legal or illegal within a specific social framework. In the context of this study focusing on the judicial sector, regulative institutions pertain to the regulations set forth by the head of the Judiciary (i.e., Chief Justice) regarding e-Justice systems.

Conversely, normative institutions encompass norms and values that categorize actions as either legitimate or illegitimate within a given social setting (Scott 2004). These encompass both social and professional norms that outline acceptable behaviors based on long-standing practices, traditions, and values. Normative institutions establish a framework that helps organizations recognize actions deemed legitimate. In this study, these institutions are associated with the norms, traditions, and practices that prevail within the socio-technical and economic environment relevant to the use of the e-Justice system.

Lastly, cognitive institutions are "taken-for-granted customs and traditions that influence social actors' sense-making and decision-making processes" (Scott 2004). The cognitive pillar thus refers to the innate concepts and frames through which meaning is produced (Dacin et al. 2002). They are intimately connected to the traditions, behaviors, and cognitive frameworks of these individuals (Grewal and Dharwadkar 2002). As concerning this research, cognitive institutions influence the patterns of thought that underline decision-making processes among participants within the e-Justice system, shaping how they perceive, interpret, and engage with the digital judicial framework. Scott (2004) highlights that although the concepts of normative, regulative, and cognitive institutions are distinctly categorized for analytical purposes, they frequently intermingle in real-world applications.

Exploring how regulative, normative, and cultural-cognitive pillars interact within Ghana's judicial sector is vital for grasping the complexities of e-Justice system adoption and usage. Regulative forces, shown through mandates from judicial leaders for digital technology use, can spur adoption but might face resistance if viewed as top-down impositions. Normative pressures, reflecting global trends towards digitalization, help shift cultural attitudes towards embracing digital practices in judicial processes, enhancing stakeholder engagement. Conversely, cultural-cognitive barriers, like resistance from those

accustomed to traditional methods, can impede effective use. The interplay between these pillars not only affects the adoption rate but also the overall effectiveness of e-Justice systems. Understanding and strategically aligning these forces is crucial for creating an environment conducive to the successful deployment and ongoing utilization of e-Justice systems in Ghana.

We cite a few studies that applied the institutional theory as their theoretical lens. Soares et al. (2021) focused on the hospitality industry in Northeastern Brazil, discovering isomorphism as hotels mimicked successful models to stay competitive. Their findings indicate that mimetic pressures were the primary drivers of technology adoption, as hotels strived to remain competitive by emulating successful peers. Although normative pressures had some impact, coercive pressures from government entities were minimal. Jan et al. (2012) investigated the adoption of e-Learning among employees, identifying that mimetic and normative pressures significantly influenced adoption, while coercive pressures did not appear to play a substantial role. These insights are particularly relevant when applied to the judicial context of our study. For instance, in Ghana's judiciary, global digitalization trends may create mimetic pressures, encouraging stakeholders to adopt e-Justice systems to meet international standards and stay competitive. Similarly, normative pressures related to professional standards and community expectations within the judiciary could either facilitate or hinder the adoption of e-Justice systems.

Integrating the findings of Soares et al. (2021) and Jan et al. (2012) with the judicial context enhances our understanding of how different forms of institutional pressures affect technology adoption in sector-specific environments. While the hospitality and corporate sectors experience internal and competitive pressures, the judicial sector may face high stakes influenced by external governmental mandates. This juxtaposition highlights the critical nature of adopting technologies in the judiciary, where the consequences of failures are more profound, necessitating a cautious and well-considered implementation strategy. By comparing these studies with the judicial sector in Ghana, we deepen our theoretical and practical understanding of how institutional theory can inform technology adoption and usage strategies in sensitive environments. This comparison not only underscores the similarities in institutional pressures across sectors but also illuminates the unique challenges and stakes in the judicial context, providing a comprehensive framework for analyzing and guiding technology adoption in the judiciary.

This study therefore moves from adoption to understanding the “regulative, normative, and cognitive” institutional influence on e-Justice usage in the judicial sector of Ghana

Research Setting and Methodology

Research Setting

The study was conducted in Ghana as a component of a broader research project. The research setting highlights the gradual integration of technology into Ghana’s judicial processes, detailing the adoption of various technological tools like voice recorders, speakers, typewriters, stenography machines, and computers across High Courts to streamline operations. Initiated in 2002, the Fast Track High Court was notably equipped with advanced technologies, significantly improving the efficiency and speed of handling cases. Although restrictions on electronic devices were observed between 2007 and 2008, subsequent years saw the acceptance of digital evidence and the utilization of laptops and projectors in courtrooms. Despite these technological advances, traditional printed legal documents remained essential in some instances. The landmark 2012 election petition case marked a significant moment in Ghana's legal history by being broadcast live on television, showcasing the evolving relationship between technology and the judicial process.

The e-Justice system is a key component of the Ghanaian Government’s initiative to improve public sector services through e-government projects. Launched in 2018, this initiative targets specifically the enhancement of judicial service delivery, focusing on reducing delays. It commenced with the High Courts situated at the Law Court Complex, Accra. Despite encountering several challenges, significant progress has been made to update the judiciary’s ICT infrastructure to better perform its functions. Collaborating with the Ministry of Communications and Digitalization, supported financially by the World Bank, this venture is a segment of the broader e-Ghana Project aimed at modernizing the Judiciary’s court automation services to augment the administration of justice in Ghana. The e-Justice system incorporates a suite of applications designed to streamline court processes, including e-case management, lawyer e-portals, e-records

management, SMS e-notifications, e-filing, public e-information services, e-payment for court dues, e-judicial decision aids, access to digital laws, and audio-video court session recordings.

Methodology

This research utilizes an interpretive qualitative single case study method within Ghana's Judicial sector. The foundation of the interpretive approach is built upon the ontological and epistemological understanding that both reality and knowledge are subjectively constructed, emerging through interactions between researchers and subjects (Orlikowski & Baroudi, 1991; Myers, 2009;). As such, interpretive research does not seek objectivity; rather, it acknowledges and incorporates subjectivity in analyzing research phenomena, their processes, and outcomes (Klein & Myers, 1999). This study, adhering to the interpretive framework, aims to delve into how e-Justice usage may affect the institutional environment. The justification for the choice of the interpretive qualitative case study method is that the subjective perspectives, experiences and interpretations of the individual stakeholders of the e-Justice system are crucial for explaining the variances in the usage of the systems. Interpretive paradigm resonates well with the purpose of this study rather than objectivity approaches like positivist and critical theory as it allows for the exploration of contextual insights that deterministic and quantitative methods might overlook. Considering the study's deductive approach, the researchers deliberated and reached a consensus on the data collection methods to ensure they were consistent with the foundational theory's concepts and the key themes highlighted in existing literature. As a result, data was sourced from a diverse source, encompassing interviews, observations, document analyses, and media content. This multi-source data collection strategy is typical of interpretive research, allowing researchers to use data from multiple sources (Walsham, 1995).

Data collection was carried out in two stages. Initially, it was essential for us to acquaint ourselves with the setting, the key stakeholders, and the e-Justice system itself. To achieve this, we employed observational methods, engaging in experiential learning and spending six months with stakeholders at the High Court in Ghana. As part of understanding the system, we reviewed project documents, reports, and user manuals. The primary data collection involved both observations and open-ended interviews, with the latter designed to allow participants the freedom to express their views while lingering within the scope of the research. To select interview participants, we used purposive sampling and snowballing techniques, ensuring a representation of the various stakeholders involved with the e-Justice system. Elaborating on the multi-source data collection strategy used in this study could enhance understanding of its contribution to the research findings. This strategy—comprising interviews, observations, document analysis, and media content—provides a comprehensive view of the institutional influences on e-Justice system adoption in Ghana's judiciary. Interviews offer qualitative insights from various stakeholders, observations add real-time usage data, document analysis reveals formal policies and media content reflects public perceptions. Together, these methods provide a multi-dimensional perspective on how regulative, normative, and cultural-cognitive forces interact, significantly enriching the analysis and ensuring a nuanced understanding of technology adoption in the judiciary.

Our interviewees encompassed a broad spectrum of roles, comprising judges, lawyers, court clerks, bailiffs, registrars, court recorders, auctioneers, filing clerks, finance officers, human resource personnel, procurement officers, auditors, litigants, and IT staff. Interview questions focused on institutional pressures and how they influenced the usage of the e-Justice system. Twenty (21) respondents were interviewed at a time convenient for them, following their agreement to participate. Semi-structured interviews were conducted (Myers and Avison, 1997) and ranged from 20 minutes to 50 minutes in duration.

Interviewees	Roles
21	Judge (3), Lawyer (3), Bailiffs (2), Registrars (1), Finance (1), Court Recorders (1), Auctioneer (1), Procurement Officer (1), Filing Clerk (1), Human Resource (1), Auditor (1), Litigants (2), Court Clerks (2), IT (1).

Table 1. Categories of Interviewees

Data Analysis

The data analysis followed a theory-driven thematic approach. After the authors gathered data from multiple sources, the next activity was to identify relevant themes about Institutional forces that may influence the use of e-Justice system in the Judicial sector of Ghana. The authors carefully read, summarized, and grouped data under its suitable theme. By leveraging the institutional theory, the authors found themes that related to normative, regulative and cognitive pressures in the usage of the e-Justice system. As (Klein and Myers 1999) posit, in relation to interpretive analysis, the objective is not to test the theory but to use it as a lens for sensitization purposes. During the analysis, follow-up interviews were undertaken to verify interim emerging findings where necessary according to the principles of the hermeneutic circle (Boell and Cecez-Kecmanovic 2014; Webb and Pollard 2006). The findings of the study are reported in the next section.

Empirical Findings

From the study and interviews conducted, the following empirical findings were deduced.

Head of Judiciary (Chief Justice Directive)

The directive of the chief justice, the head of judiciary was identified as a regulatory force in the use of the e-justice system. According to several of the stakeholders interviewed, the e-Justice system had not been given much attention in terms of usage before COVID-19. Even though efforts to institutionalize the e-Justice system were way before the break out of the COVID-19 pandemic in March 2020, the process had so much been slow-paced. After the pandemic, the Chief Justice aiming to hasten the usage of the e-Justice system gave a directive and regulation to the judiciary by which he emphasized the need for them to use the digital technologies in the judicial process (Ghana 2024). This directive was made to remove case backlogs, deal with cases on time, and by so doing become more effective and efficient. One Judge described the regulation as follows:

“I found the timing crucial and necessary due to the numerous case backlogs caused by the COVID-19 lockdown, which had delayed all scheduled cases. Following the directive from the Chief Justice, I began utilizing the video conferencing feature of the integrated e-Justice system to conduct virtual court hearings for my assigned cases." ***(Judge)***

More so, for the legal vacation period in August and September 2022, the Chief Justice directed that all vacation courts within the Greater Accra Region operate through virtual sessions using Microsoft Teams. Participants could access the virtual vacation courts through a link provided on the Judicial Service of Ghana's official website. Finally, one lawyer disclosed that for the Chief Justice to ensure that they comply with the regulation to use the e-Justice system, their renewal of their license as lawyers is tied to its usage for their duties. The lawyer said:

“I have been utilizing the e-Justice system for an extended period and have greatly appreciated the convenience it offers, allowing me to file cases from home, receive invoices, and make payments via mobile money. Additionally, given that the use of this system is linked to the renewal of our legal licenses, I am compelled to continue using it” ***(Lawyer)***

E-Justice systems gaining popularity across the globe

E-Justice system deployment in the judicial sector becoming popular across the globe was revealed as a normative force in the use of the e-Justice system. The data analysis also shows the normative use of the e-Justice system in Ghana. The researchers found out that the stakeholders use the system because they believe it has become a norm in many settings both in the developed and developing worlds. Some even acknowledged that they were aware of the remarkable benefits that using the e-Justice system has brought developed nations and so they were receptive to using the technologies. Three interviewees said:

"Technology is now central to modern life, especially in the 21st century. In Europe and across the Western world, e-Justice systems have transformed the litigation process so profoundly that they've become the standard. We need to adopt this norm as well." ***(Court Clerk)***

"I connected with friends who are judges on LinkedIn last year, and communicating with them, I realized they use systems deployed in their context most times in dispensing their duties as judges, especially after the COVID-19 pandemic. So, as a judge here in Ghana, I realized that if I use the system, it will help me in numerous ways." ***(Judge)***

""Well, technology, as you and I know, has pervaded all sectors, automating manual tasks and making organisations effective. So, for me, it is a good thing that the judicial sector of Ghana has also made this remarkable stride to use technology because it is the new norm for every organization. As IT people in this department, we will make sure the system is always up-to-date to help in the running of the judicial sector." ***(IT Support Officer)***

Unwillingness to Use

Cognitive institutions have been identified to influence the usage of the e-Justice system. It is worth noting that, some stakeholders' cognitive thinking also affects their decision in the usage of the e-Justice system. According to a senior judge well advanced in age, he does not use the e-Justice system. He told the authors that by using the system, some of the standing orders and practices done in physical courtrooms become almost impossible to enforce. He expressed his mind as follows:

"I don't use it in adjudicating cases. Some customs are observed during litigation. For example, everyone in court rises when the judge enters, by way of revering that the judge is institutionalised. But you know, by using the system for virtual hearings, this, among others, becomes unimplementable." ***(Judge)***

Analysis and Discussions

Scholarly research on e-Justice within developing nations has predominantly concentrated on the processes of its adoption and implementation (Abd Ghadas and Mohd Ariffin 2019; Aboelazm 2022; Apriyanti 2021; Kumar Pandey 2023; Martínez et al. 2023; Rosa et al. 2013; Salah-Eddine 2022) with less attention on the institutional effect influencing the use of the systems. This study extends the literature from just the adoption and implementation of e-Justice systems to investigate the institutional factors that influence its actual usage in a developing country context, precisely, Ghana. Therefore, the purpose of this study has been to comprehend how institutional forces —cultural-cognitive, regulative, normative— influence e-Justice system usage in the judicial sector of Ghana.

The findings make it evident that regulative institutions affect e-Justice system usage. Regulative force entails the ability of entities to be governed by rules or laws. Illuminating this section, the data shows that the Chief Justice of Ghana's directive compelled the stakeholders to use the e-Justice system. Even though the usage of digital technologies in judicial processes had been slow, after the regulation by the Chief Justice the e-Justice system usage saw some considerable progress.

More so, the study showed the normative use of the e-Justice system. Some of the stakeholders recount that, in the developed world, the judicial sector has been inundated with digital technologies which has made their justice system transparent, and robust making them efficient in their litigations. They opined that they use the system not just because it is mandated but because it is the norm and professional standard presently. The study also showed cognitive institutions affect the e-Justice system usage.

Cognitive pillars involve the customs and traditions that govern the process of sense-making, influencing how social actors make decisions, which in this study are the stakeholders of e-Justice. The study findings showed that some stakeholders own cognitive institutions makes them unwilling to use the e-Justice system. In sum, the findings suggest that e-Justice systems discourse should move away from adoption and implementation and delve into institutional forces and their influence on e-Justice usage most especially in developing countries.

Conclusion

This study seeks to answer how the regulative, normative, and cultural cognitive institutional pillars shape e-Justice usage in Ghana's judicial sector. The study identified that institutional forces (regulative, normative, and cognitive) do influence e-Justice system usage. Regulations by heads of the judiciary obliged the stakeholders to use the system. Again, the findings showed that some stakeholders were ready to use

the e-Justice system because it is the norm currently across both the developing and developed world. Finally, the findings also disclosed that some stakeholders' cognitive frame affected the decision not to use the system. The originality of this work is enshrined in understanding how institutional pillars influence e-Justice system usage in a developing country, precisely, Ghana.

This study has contributed to theory, research, and policy. For research, this study suggests that investigating the e-Justice phenomenon should move from the mere adoption and diffusion to ascertain different institutional forces that can unearth the dynamics that pertain to its usage. For theory, this is the foremost study to use institutional theory as a lens in studying the e-Justice system usage in the developing country context of Ghana. For policy, heads of the judiciary should have a mechanism in place that will enable them to figure out those who defy the regulations and mate out disciplinary measures against them. Again, heads of the Judiciary could give some incentives in terms of promotion, and rewards among others to those who use the system. This study comes with certain limitations. First and foremost, the study is a single case research. Future studies can adopt a multiple-case study scenario and do a comparative analysis of cases even from other jurisdictions. Finally, the institutional theory is adopted as the lens to investigate institutional pressures that influence the use of the e-Justice system. Future studies could also explore other services like e-passports, e-education, and e-health.

REFERENCES

Abd Ghadas, Z. A., and Mohd Ariffin, R. A. 2019. "E-Court System in the Civil and Shariah Courts: Malaysia Perspectives," *Proceedings of the International Conference on Islamic Civilization and Technology Management*, p. 24.

Aboelazm, K. 2022. "The Role of Digital Transformation in Improving the Judicial System in the Egyptian Council of State," *Journal of Law and Emerging Technologies* (2:1), pp. 11–50.

Aksom, H., and Tymchenko, I. 2020. "How Institutional Theories Explain and Fail to Explain Organizations," *Journal of Organizational Change Management* (33:7), pp. 1223–1252. (https://doi.org/10.1108/JOCM-05-2019-0130).

Apriyanti, I. 2021. "Emerging Courts and Judicial Integrity: Challenges in Digital Transformation of the Indonesia Constitutional Court," Jakarta. (https://en.mkri.id/icjf/public/upload/papers/paper presentation R.A. Indah Apriyanti for Short course 2023.pdf).

Avgerou, C. 2000. "IT and Organizational Change: An Institutionalist Perspective," *Information Technology & People* (13:4), pp. 234–262. (https://doi.org/10.1108/09593840010359464).

Bawole, J. N., and Langnel, Z. 2022. "Administrative Reforms in the Ghanaian Public Services for Government Business Continuity During the COVID-19 Crisis," *Public Organization Review*, Springer. (https://doi.org/10.1007/s11115-022-00687-w).

Bharadwaj, A., Sawy, O. A. El, Pavlou, P. A., and Venkatraman, N. 2013. "Digital Business Strategy: Toward a Next Generation of Insights," *MIS Quarterly* (37:2), pp. 471–482. (https://www.jstor.org/stable/pdf/43825919.pdf?refreqid=excelsior%3Af70f7a2f574f93c4cfabf2877c4c2079).

Bloomberg, J. 2018. *Digitization, Digitalization, And Digital Transformation: Confuse Them At Your Peril*, pp. 2–7.

Bygstad, B., Egil, Ø., Ludvigsen, S., and Dæhlen, M. 2022. *From Dual Digitalization to Digital Learning Space: Exploring the Digital Transformation of Higher Education*, (182:February). (https://doi.org/10.1016/j.compedu.2022.104463).

Dacin, M. T., Goodstein, J., and Scott, W. R. 2002. "Institutional Theory and Institutional Change: Introduction to the Special Research Forum," *Academy of Management Journal* (45:1), pp. 45–57. (https://doi.org/10.2307/3069284).

Danyliuk, M., Dmytryshyn, M., and Goran, T. 2021. "Digitisation of Ukraine in Terms of Public Electronic Services' Distribution," *Scientific Horizons* (24:7), pp. 90–99. (https://doi.org/10.48077/scihor.24(7).2021.90-99).

Đilas, A., and Kuzmanović, D. 2023. *Digital Transformation in the Judiciary*, pp. 50–56. (https://etakse.sud.rs/,).

DiMaggio, P. J., and Powell, W. W. 1983. "The Iron Cage Revisited: Institutional Isomorphism and Collective Rationality in Organizational Fields Author (s): Paul J . DiMaggio and Walter W. Powell Published by: American Sociological Association Stable URL: Http://Www.Jstor.Org/Stable/2095101," *American Sociological Review* (48:2), pp. 147–160.

Drossos, D., Lekakos, G., Doukidis, G. I., Tsatsa, N., and Coursaris, C. K. 2018. "Public-Sector Digitization: An Analytical Cost-Effective Framework," in *MCIS 2018 Proceedings. 38*. (https://aisel.aisnet.org/mcis2018/38).

Fitzgerald, M., Kruschwitz, N., Bonnet, D., and Welch, M. 2013. "Embracing Digital Technology. A New Strategic Imperative." (http://sloanreview.mit.edu/faq/).

Grewal, R., and Dharwadkar, R. 2002. "The Role of the Institutional Environment in Marketing Channels," *Journal of Marketing* (66:3), pp. 82–97. (https://doi.org/10.1509/jmkg.66.3.82.18504).

Jan, P.-T., Lu, H.-P., and Chou, T.-C. 2012. "The Adoption of E--Learning: An Institutional Theory Perspective," *Journal of Educational Technology* (11:3), pp. 55–61

Jneid, M., Saleh, I., and Fakhoury, R. 2019. "Digital Transformation in Justice: Discussion of Challenges and a Conceptual Model for e-Justice Success," *Proceedings of the European Conference on E-Government, ECEG* (2019-October), pp. 35–42. (https://doi.org/10.34190/ECDG.19.051).

Judiciary of Ghana 2024. "CTS Requires Full Stakeholder Commitment to Achieve Success – CJ." (https://www.judicial.gov.gh/index.php/publications/news-publications/js-latest-news/item/425-cts-requires-full-stakeholder-commitment-to-achieve-success-cj, accessed February 28, 2024)

Klein, H. K., and Myers, M. D. 1999. "A Set of Principles for Conducting and Evaluating Interpretive Field Studies in Information Systems," *MIS Quarterly: Management Information Systems* (23:1), pp. 67–94. (https://doi.org/10.2307/249410).

Klein, H. K., and Myers, M. D. 2001. "A Classification Scheme for Interpretive Research in Information Systems," in *Qualitative Research in IS: Issues and Trends*, IGI Global, pp. 218–239.

Kondori, N. F., and Rouhani, S. 2021. "Presenting a Conceptual Framework for Digital Judicial Transformation for Digital Governance," *Journal of Public Administration* (13:3), pp. 593–620. (https://doi.org/10.22059/JIPA.2021.317608.2891).

Kshetri, N. 2007. "Institutional Changes Affecting Entrepreneurship in China," *Journal of Developmental Entrepreneurship* (12:4), pp. 415–432.

Kumar Pandey, A. 2023. "Access To Justice Through Virtual Courts In Current Indian Scenario," *South Asian Law & Economics Review* (08), The Law Brigade Publishers, pp. 33–44. (https://doi.org/10.55662/saler.2023.801).

Kuusisto, M. 2017. *Organizational Effects of Digitalization: A Literature Review*, (20:3), pp. 341–362.

Liere-Netheler, K., Packmohr, S., and Vogelsang, K. 2018. "Drivers of Digital Transformation in Manufacturing," *Proceedings of the Annual Hawaii International Conference on System Sciences* (2018-Janua), pp. 3926–3935. (https://doi.org/10.24251/hicss.2018.493).

Loo, J., and Findlay, M. 2022. "Digitised Justice: The New Two Tiers?," *Criminal Law Forum* (33:1). (https://doi.org/10.1007/s10609-022-09431-x).

Marta, G. C., and Giulia, G. 2023. "Algorithms, Rule of Law, And The Future of Justice: Implications in the Estonian Justice System

Martínez, A., Ramiro, M., Burgos, L., Paulina, G., López, P., Alfredo, J., Cevallos, G., and Patricio, D. 2023. "Ecuador and Digital Justice in Time of the Covid-19 Pandemic," *Baltic Journal of Law and Politics* (16:3), pp. 1395–1408. (https://doi.org/10.2478/bjlp-2023-00000113).

Murphy, W. F., and Pritchett, C. H. 1986. *Courts, Judges, and Politics - An Introduction to the Judicial Process*, (4th ed.). (https://www.ojp.gov/ncjrs/virtual-library/abstracts/courts-judges-and-politics-introduction-judicial-process-fourth).

Okano, M. T., de Castro Lobo dos Santos, H., Ursini, E. L., Honorato, W., and Ribeiro, R. B. 2021. "Digital Transformation as an Agent of Change in a Pharmaceutical Industry from the Perspective of Dynamic Capabilities," in *2021 IEEE 11th Annual Computing and Communication Workshop and Conference (CCWC)*, IEEE, January 27, pp. 0896–0902. (https://doi.org/10.1109/CCWC51732.2021.9375838).

Parida, V., Sjödin, D., and Reim, W. 2019. *Reviewing Literature on Digitalization , Business Model Innovation , and Sustainable Industry : Past Achievements and Future Promises*. (https://doi.org/10.3390/su11020391).

Preko, M., Boateng, R., and Effah, J. 2019. "Healthcare Digitalisation in Ghana: An Assessment Based on the Realist Social Theory," *25th Americas Conference on Information Systems, AMCIS 2019*, pp. 1–5.

Reis, J., Amorim, M., Melão, N., Cohen, Y., and Rodrigues, M. 2020. "Digitalization: A Literature Review and Research Agenda," *Springer Nature Switzerland AG*. (https://doi.org/10.10017/978-3-030-43616-2_47).

Rosa, J., Teixeira, C., and Sousa Pinto, J. 2013. "Risk Factors in E-Justice Information Systems," *Government Information Quarterly* (30:3), The Authors, pp. 241–256. (https://doi.org/10.1016/j.giq.2013.02.002).

Salah-Eddine, L. 2022. "Digital Transformation in the Moroccan Public Sector: Literature Review, State of the Art and Perspectives.," *Revue Francaise d'Economie et de Gestion* (3), pp. 86–103. (www.revuefreg.fr).

Sandoval-Almazan, R., and Gil-Garcia, J. R. 2020. "Understanding E-Justice and Open Justice Through the Assessment of Judicial Websites: Toward a Conceptual Framework," *Social Science Computer Review* (38:3), pp. 334–353. (https://doi.org/10.1177/0894439318785957).

Scott, W. R. 2004. "Institutional Theory: Contributing to a Theoritical Research Program," in *Great Minds in Management: The Process of Theory Development*, Oxford UK: Oxford University Press, pp. 1–96. (https://www.researchgate.net/publication/265348080).

Shevchenko, Y. 2022. *The Digitization of Administrative Justice in Ukraine*, pp. 288–299. (http://repository.mruni.eu/handle/007/18324).

Soares, A. L. V., Mendes-Filho, L., and Gretzel, U. 2021. "Technology Adoption in Hotels: Applying Institutional Theory to Tourism," *Tourism Review* (76:3), pp. 669–680. (https://doi.org/10.1108/TR-05-2019-0153)

Tilson, D., Sørensen, C., and Lyytinen, K. 2010. *Digital Infrastructures: The Missing IS Research Agenda*, (December 2013). (https://doi.org/10.1287/isre.1100.0318).

Tilson, D., Sørensen, C., and Lyytinen, K. 2012. *Change and Control Paradoxes in Mobile Infrastructure Innovation: The Android and IOS Mobile Operating Systems Cases*, (2012: January), IEEE, pp. 1324–1333. (https://doi.org/10.1109/HICSS.2012.149).

Tsindeliani, I. A., Proshunin, M. M., and Sadovskaya, T. D. 2022. *Digital Transformation of the Banking System in the Context of Sustainable Development*, (25:1), pp. 165–180. (https://doi.org/10.1108/JMLC-02-2021-0011).

Tungpantong, C., Nilsook, P., and Wannapiroon, P. 2021. "A Conceptual Framework of Factors for Information Systems Success to Digital Transformation in Higher Education Institutions," in *2021 9th International Conference on Information and Education Technology, ICIET 2021*, Institute of Electrical and Electronics Engineers Inc., March 27, pp. 57–62. (https://doi.org/10.1109/ICIET51873.2021.9419596).

Walsham, G. 1995. "Interpretive Case Studies in IS Research: Nature and Method," *Eur. J. Inf. Systs* (4), pp. 74–81.

Yoo, Y. 2010. "Digitalization and Innovation," Tokyo, Japan.

SPEAKER

Second African Digital Humanism Summer School, Accra, Ghana, 31 August – 05 September 2026

Julia Neidhardt

Julia Neidhardt is an Assistant Professor at the Data Science Research Unit at TU Wien, Vienna, Austria, where she leads the Christian Doppler Lab for Recommender Systems. She has held the UNESCO Co-Chair on Digital Humanism since 2023. She has a background in mathematics and computer science and has been a guest researcher at the Austrian Academy of Sciences, Northwestern University, and the University of Geneva. Her research lies at the intersection of user modeling, recommender systems, and Digital Humanism, with applications in domains such as news, e-commerce, and tourism. Her recent work focuses on generative conversational recommender systems and the impact of digitalisation on linguistic diversity. She is actively involved in the international research community through editorial, program committee, and organizational roles, and serves on the board of the TU Wien Center for AI and Machine Learning (CAIML).

Lecture 6: *Recommender Systems and Their Ethics*

Reader article:

Peter Knees, Julia Neidhardt, and Irina Nalis: Recommender Systems: Techniques, Effects, and Measures Toward Pluralism and Fairness. In H. Werthner, C. Ghezzi, J. Kramer, J. Nida-Rümelin, B. Nuseibeh, E. Prem, A. Stanger (Eds.): Introduction to Digital Humanism – A Textbook, pp. 417-434. Springer Nature, Cham, Switzerland, 2024. https://doi.org/10.1007/978-3-031-45304-5_27

Recommender Systems: Techniques, Effects, and Measures Toward Pluralism and Fairness

Peter Knees, Julia Neidhardt, and Irina Nalis

Abstract Recommender systems are widely used in various applications, such as online shopping, social media, and news personalization. They can help systems by delivering only the most relevant and promising information to their users and help people by mitigating information overload. At the same time, algorithmic recommender systems are a new form of gatekeeper that preselects and controls the information being presented and actively shapes users' choices and behavior. This becomes a crucial aspect, as, if unaddressed and not safeguarded, these systems are susceptible to perpetuate and even amplify existing biases, including unwanted societal biases, leading to unfair and discriminatory outcomes. In this chapter, we briefly introduce recommender systems, their basic mechanisms, and their importance in various applications. We show how their outcomes and performance are assessed and discuss approaches to addressing pluralism and fairness in recommender systems. Finally, we highlight recently emerging directions within recommender systems research, pointing out opportunities for digital humanism to contribute interdisciplinary expertise.

1 Introduction

Recommender systems (RSs) are software tools and techniques that use data from users to suggest items they will probably like. These suggestions cover a wide range of items including books, news articles, music, as well as more complex products such as tourist offers, job vacancies, or financial loans. Today, recommender systems are widely adopted. Personalization techniques are used by all major online platforms such as Google, Amazon, Facebook, YouTube, Netflix, Spotify, Booking.

P. Knees (✉)
Faculty of Informatics, TU Wien, Vienna, Austria
e-mail: peter.knees@tuwien.ac.at

J. Neidhardt · I. Nalis
Christian Doppler Lab for Recommender Systems, TU Wien, Vienna, Austria
e-mail: julia.neidhardt@tuwien.ac.at; irina.nalis-neuner@tuwien.ac.at

H. Werthner et al. (eds.), *Introduction to Digital Humanism*,
https://doi.org/10.1007/978-3-031-45304-5_27

com, LinkedIn, and many others to tailor their services to the specific preferences and needs of individual users. It can be argued that recommender systems provide a backbone to the modern, industrialized Web, as they facilitate—as well as steer—access to the world's digital content. As such, they have become a new, albeit more subtle form of gatekeeper for information, culture, and other resources, built on technological opportunities and interests of those operating them.

The implications of incorporating automatic recommender systems to structure access to information and digital goods in virtually all large-scale Web services and platforms are wide-reaching and have led to increased interest from the general public. While initially being welcomed as another commodity of the digital world that effortlessly identifies matching content, limitations and frustrations have soon led to a more critical view of their effects. They become of particular interest when they have the potential to affect society and democratic processes, such as "filter bubbles" (Pariser, 2011).

From the perspective of digital humanism, technology that threatens democracy and leads to isolation of individuals must be redesigned and shaped in accordance with human values (Werthner et al., 2019). More specifically, artificial intelligence (AI) and automated decision-making systems such as recommender systems are often resulting in black-box models for which decision processes remain intransparent, bias unknown, and results unfair (Werthner et al., 2023). As such, we need to understand the principles of recommender systems and analyze strategies to overcome a situation where the mechanisms of the applied method and/or the characteristics of the underlying data dictate undesired and unfair outcomes.

In this contribution, we focus on two desiderata for recommender systems with possible broader societal implications: diversity, as a proxy for pluralism, and fairness. Specifically, we outline how recommender systems can be turned from a threat to pluralism and fairness to an instrument for promoting them. As always, the situation is complex, and no single (technical) solution can fully remedy the undesired effects. After describing the main concepts, methods, and practices in recommender systems research, we discuss the concepts of diversity and fairness in the context of filter bubbles. This is followed by a discussion of optimization goals beyond accuracy, such as diversity, and fairness in more detail. We then investigate methods and research directions for promoting both concepts, diversity and fairness. Finally, we touch on the emerging field of moral and human-centered recommender systems. We include examples to illustrate the concepts and methods discussed before summing the discussion up by presenting the main take-away messages. This discussion is continued and deepened in the following contribution of this chapter, where topics of bias, balancing of various and diverse interests (see chapter of Baeza-Yates and Murgai), and automatic content moderation (see chapter of Prem and Krenn) are addressed.

In the following, we investigate the topics of recommender systems and optimization goals beyond accuracy, such as diversity and fairness in more detail, and highlight current research directions.

2 Recommender Systems: Concepts and Practices

RSs are software applications that use a variety of data sources, including users' past behavior and preferences, to suggest items (e.g., goods, articles, content) to a user that he/she is likely to find interesting (Ricci et al., 2022). The overall schema of a RS is illustrated in Fig. 1. To provide personalized suggestions, the RS needs to have knowledge about the items as well as knowledge about the users. With respect to the items, this knowledge can include textual descriptions, keywords, genres, product categories, release date, or price. With respect to the users, demographic data such as age and gender; data about a user's past behavior such as previous purchases, clicks, or ratings; or a user's online social network are commonly used by RSs. Importantly, a relationship between the two sides, items and users, has to be established by the RS so that it knows which items are possibly enjoyed by a specific user. This relationship is typically established using previous purchases, clicks, ratings, or other behavioral data. There are several fundamental recommendation approaches that are traditionally distinguished, which all exploit different aspects to determine those items to be suggested to a user (Burke, 2007; Ricci et al., 2022): In the content-based approach, items are recommended that have similar attributes to those that the user previously liked (e.g., same category). With collaborative filtering, items liked by users with similar preferences are considered important (e.g., "users who bought this also bought that"). Demographic systems recommend items based on user demographics (e.g., items are recommended that are popular in a specific age group). Knowledge-based approaches make recommendations based on domain knowledge about how specific item properties match user preferences (e.g., a travel recommender system that leverages domain knowledge about various travel destinations and their properties). Community-based approaches recommend items liked by the

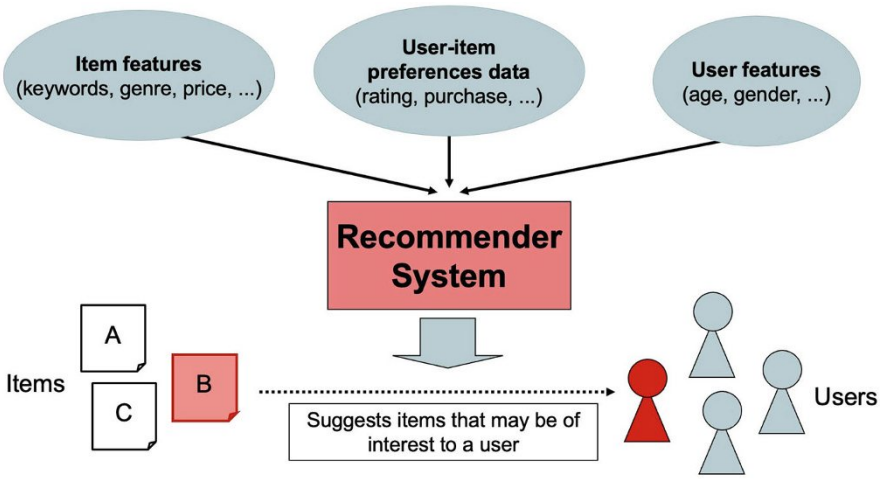


Fig. 1 Overall structure of a RS

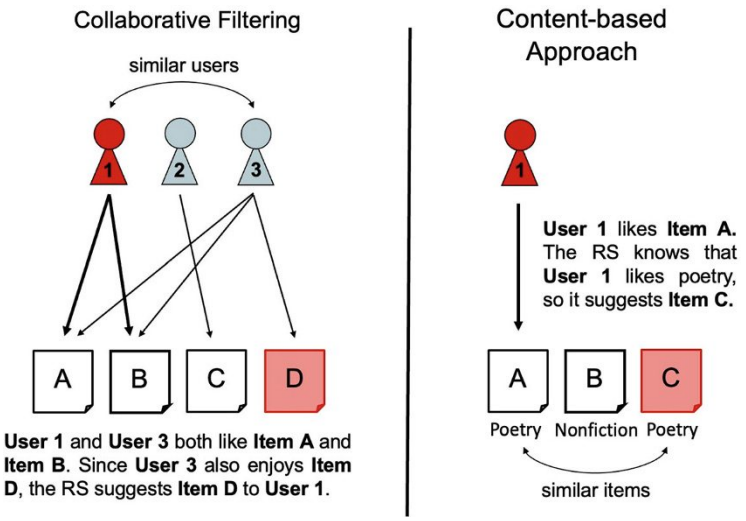


Fig. 2 Collaborative filtering vs. content-based approach

user's friends, often within an online social network. Hybrid recommender systems combine different recommendation techniques to make more accurate and personalized recommendations. Since recommender systems aim to offer personalized suggestions, all the techniques mentioned rely on knowledge about the user. Therefore, every RS needs to include a user model or user profile where this knowledge is accumulated and stored (Jannach et al., 2010). However, this dependency on user data gives rise to significant concerns regarding privacy and misuse.

RS can face a challenge when they encounter new users since there may not be enough data to build a user model. This situation is called "cold-start problem," and the best way to address it depends on the specific use case. The same is true for items that are new to the system. In Fig. 2, the two most common RS approaches, i.e., collaborative filtering and the content-based approach, are conceptually shown. Collaborative filtering encounters the cold-start problem when there is a lack of user-item interaction data, making it challenging to identify similar users. For new users, the recommender system may suggest popular items initially. Similarly, for new items, the system faces difficulties in making recommendations until some users interact with them. In contrast, content-based approaches overcome the cold-start problem for new items by relying on the item's inherent characteristics or features. The system can recommend the item to users who have shown interest in similar items, even if the item has not been previously interacted with. Additionally, for new users, the system can provide recommendations based, for example, on user-provided preferences during the onboarding process.

In the last few years, deep learning architectures have been increasingly used in recommender systems, particularly for capturing various patterns and dealing with high complexity (Zhang et al., 2019). Large language models have also emerged as powerful tools within recommender systems very recently (Liu et al., 2023).

Traditionally, recommendation approaches have focused on predicting how a given user would rate certain items. These approaches are typically tested through so-called offline evaluation, where actual ratings are withheld and used for forecast assessment. The better a method can accurately predict the withheld ratings, the more successful it is. This evaluation approach has significantly advanced the field of recommendation systems. However, there are limitations to this approach such as the absence of real-time feedback, the limited availability of contextual information, and the inability to directly measure user satisfaction (Jannach et al., 2016). To address these limitations, offline evaluation is often supplemented with online evaluation, user studies, and A/B testing, which represent a more realistic and dynamic assessment of recommender systems in real-life settings.

Relying solely on accuracy measurements and a lack of diversity in tailored content consumption can introduce bias and lead to filter bubbles, echo chambers, and related phenomena (Stray et al., 2022). Specifically, users may become trapped in echo chambers and filter bubbles when only considering users' preexisting likes and interests in order to produce the most accurate recommendations, which may lead to a lack of "media pluralism" or the exposure to and consumption of a variety of information, ideas, and viewpoints (Vermeulen, 2022). Correspondingly, there is a growing recognition that the quality of a recommender system extends beyond accuracy measurements only.

3 Recommender Systems as a Threat to Pluralism and Fairness?

An often-discussed effect of automatic information filters such as recommender systems is a loss of diversity in the presented options, due to emphasizing similarity to previous choices. This has been branded and popularized by Pariser (2011) as "filter bubbles." For instance, when consuming social media, and showing interest in posts and articles dealing with, say, migration, recommender systems can pick up this signal and increasingly suggest content dealing with migration. This might lead to overrepresentation of the topic in the shown posts and oust other topics potentially of interest. As such, the topic of migration, despite originally being of (temporal) interest, eventually disproportionally occupies space in the contents recommended and continues to draw the user's attention. Moreover, the recommender system might increasingly present posts from the authors of the consumed posts, i.e., the providers, ousting other authors and leading to a loss of diversity in sources.

Pariser (2011) argues that algorithmic information filters and personalized services are directly connected to individualization and intellectual isolation, ultimately leading to polarization and social fragmentation, posing a threat to democratic societies. Several works have subsequently investigated this connection (e.g., Nguyen et al., 2014; Aridor et al., 2020) and found inconclusive results regarding users' behavior upon usage of recommender systems and its impact on the diversity

of items consumed over time. As a consequence, Dahlgren (2021) suggests a further differentiation between *technological filter bubbles* and their consequences that manifest in *societal filter bubbles*. To investigate the former, Michiels et al. (2022) provide an operational definition of technological filter bubbles as "a decrease in the diversity of a user's recommendations over time, in any dimension of diversity, resulting from the choices made by different recommendation stakeholders." Correspondingly, the personalization-polarization hypothesis assumes that these filter bubbles influence the division of large crowds (and thus also of society) into individual groups due to their strongly divergent opinions (Keijzer & Mäs, 2022). The importance of the concept of diversity on the technical side is linked to the societal relevance, as, e.g., stated by Helberger et al. (2018): "As one of the central communication policy goals, diversity refers to the idea that in a democratic society informed citizens collect information about the world from a diverse mix of sources with different viewpoints so that they can make balanced and well-considered decisions."

The definition by Michiels et al. (2022) further highlights the aspect of different recommendation stakeholders (Abdollahpouri & Burke, 2022). While diversity is an important mechanism to avoid one-sidedness with regard to topics and/or sources, for a recommender system, different interests are competing and ultimately might also conflict with the goal of diversity. The typical stakeholders to consider in a recommender system are the consumers (who are typically referred to as the "users"), the providers of items, and the system (or service provider) itself. They all want to ensure to not be treated in an unfair manner and optimize their gain or utility. For instance, users might be treated unfairly if the quality of service they receive depends on individual traits, if these relate to sensitive attributes such as race or gender. Item providers might be treated unfairly if they are deprived of exposure to users, for instance, by not being recommended. The system has the task of maintaining fairness toward all different stakeholders (or at least plausibly argue for it) while maximizing utility, e.g., by recommending items that are most profitable or otherwise beneficial for the system. A typical example to showcase fairness in recommendation toward multiple stakeholders are job recommendations, as performed on business-oriented social media platforms such as LinkedIn or Xing (Ekstrand et al., 2022). In addition to country-specific regulations that might also play a role, matching candidates with job offers is an inherently multi-sided fairness problem. In this scenario, job seekers and employers and both consumers and providers alike, always with the goal to obtain recommendations with the highest utility, raising questions such as: Are recommendations of job opportunities distributed fairly across users? Are job candidate fit scores fair, or do they under- or overscore certain candidates? Do users have a fair opportunity to appear in result lists when recruiters are looking for candidates for a job opening? Are employers in protected groups (e.g., minority-owned businesses) having their jobs fairly promoted to qualified candidates?

Besides fair distribution of opportunities, questions of bias with regard to certain candidates or employers, esp. in protected groups, arise. Beyond this simplified view of competing interests within a recommender system, there are potentially many

more stakeholders to consider. For instance, in music streaming, there are multiple types of providers, e.g., the composers, the record labels, or the rights owners; food order platforms add delivery drivers as stakeholders; etc.

4 Beyond Accuracy: Diversity, Novelty, and Serendipity

The "beyond-accuracy" paradigm in recommender system research has been sparked by users scrolling endlessly through items they are already familiar with or that are too similar to their current preferences. This field of study investigates different evaluation measures to improve the value and caliber of recommendations (Smets et al., 2022). Other aspects, such as diversity, serendipity, novelty, and coverage, are being more and more considered in evaluation. These concepts are briefly characterized in the following (Kaminskas & Bridge, 2016; Castells et al., 2022).

Diversity in recommender systems means including a range of different items in the recommendations for users. The goal is to provide a broad selection of items that cover various categories or genres. When the recommender system offers a diverse list of recommendations, users get to see a wide range of options. This allows them to explore and discover new items, which helps them to expand their horizons and ideally improves their overall experience. Content-based approaches (see Sect. 2) often lack diversity because they focus on recommending items that are similar with, e.g., genre. Collaborative filtering can have higher diversity than content-based approaches because it considers the preferences of other users, which can vary widely and thus expose the user to a wider variety of items. An area that attracts a growing number of studies is the domain of news recommendations. Users may not be exposed to opposing viewpoints if tailored news recommendations lack diversity (Stray et al. 2022). A news recommender must find a balance between remaining relevant to users' interests and delivering enough diversity, such as exposing users to new topics and categories, to maintain their interest. The deep neural network presented by Raza and Ding (2020) satisfies the user's requirement for information on subjects in which they have previously expressed interest while going above and beyond accuracy metrics. With an emphasis on the effectiveness of news diversity and confidence in news recommenders, Lee and Lee (2022) investigated the function of perceived personalization and diversity in services. They investigated the effects of perceived personalization and news diversity on users' inclinations to stick around and found that diversity had a positive effect on user satisfaction and continuance intention. From the perspective of the interplay of news diversity and democracy, Helberger et al. (2019) highlight the importance of perspective diversity for well-informed citizens of a democratic society. Furthermore, they underline that interests of the users (autonomy, privacy, and accuracy) need to be considered and balanced against the power and opportunities data and algorithms have to offer—herein lies a great challenge for the design of recommender systems. Another challenge for RS

lies in finding a balance between the most accurate and simultaneously diversified recommendations for the user (Möller et al., 2018; Ribeiro et al., 2015).

Occasionally used in connection with measuring diversity is *coverage*. Coverage refers to the proportion of items within the system that the recommender system can recommend to users. A high coverage indicates that the recommender system can suggest items from different genres, topics, or domains, accommodating the varied tastes of its user base.

The concept of *novelty* refers to the degree of uniqueness or freshness of the recommended items. The goal is to suggest items that the user is not familiar or has not seen before. Novel recommendations aim to introduce users to new and unexpected items, encouraging them to explore and avoid repeating previously consumed items.

Serendipity refers to the element of surprise or unexpectedness in the recommendations. It aims to suggest items that go beyond a user's explicit preferences or direct expectations. Serendipitous recommendations should surprise users by presenting them with items they did not anticipate but still find enjoyable or valuable. Serendipity has been examined regarding its potential to reduce popularity bias and boost the utility of recommendations by facilitating better discoverability. However, designing serendipity is challenging, as it requires balancing surprises as well as relevance. One line of research that combines the necessity to provide users with surprising and yet relevant items is presented by Björneborn (2017) and has seen operationalization in recent attempts to design recommender systems beyond the algorithm (Smets et al. 2022). Björneborn (2017) identifies three key affordances for serendipity: diversifiability, traversability, and sensoriability. These affordances are linked to three personal serendipity factors: curiosity, mobility, and sensitivity. Diversifiability relates to curiosity and includes factors such as interest, playfulness, and inclusiveness. Traversability is associated with mobility and encompasses searching, immersion, and exploration. Sensoriability is linked to sensitivity and involves stumbling upon, attention, surprise, and experiential aspects. The essential components of human interactions with information environments are covered by these affordances. A quintessential understanding that can be derived from this operationalization is that environments can be designed in ways that cultivate serendipitous encounters, whereas serendipity itself cannot be designed for (Smets, 2023; Björneborn, 2017).

5 Fairness

As we have seen, beyond-accuracy measures attempt to introduce other aspects besides accurately re-predicting historic interactions to evaluate recommender systems. Still, these measures focus only on the items that are recommended. However, the bigger context of who is affected in which way by the recommendations given by a system and whether the results or mechanisms underlying this are "fair" has become an increasingly important factor for evaluating recommender systems

(Ekstrand et al., 2022). In recent years, the fairness of machine learning has gained significant attention in discussions about machine learning systems. Fairness in classification and scoring or ranking tasks has been extensively studied (Chouldechova & Roth, 2020). Here concepts like individual fairness and group fairness are typically investigated. Individual fairness aims to treat similar individuals similarly, ensuring comparable decisions for those with similar abilities. Group fairness examines how the system behaves concerning group membership or identities, addressing discriminatory behaviors and outcomes. Ekstrand et al. (2022, p. 682) list the following fundamental concepts in terms of fairness definitions, harm, and motivations for fairness.

Definitions

- *Individual fairness: Similar individuals have similar experience.*
- *Group fairness: Different groups have similar experiences.*
  - *Sensitive attribute: Attribute identifying group membership.*
  - *Disparate treatment: Groups explicitly treated differently.*
  - *Disparate impact: Groups receive outcomes at different rates.*
  - *Disparate mistreatment: Groups receive erroneous (adverse) effects at different rates.*

Harm

- *Distributional harm: Harm caused by (unfair) distribution of resources or outcomes.*
- *Representational harm: Harm caused by inaccurate internal or external representation.*

Motivations

- *Anti-classification: Protected attributes should not play a role in decisions.*
- *Anti-subordination: Decision process should actively work to undo past harm.*

With regard to the motivations for fairness, the most commonly discussed and addressed aspect in technology-oriented works is that of anti-classification, i.e., to prevent harm before occurring. The concept of anti-subordination, i.e., addressing past harm and therefore introducing current "unfairness" in order to support historically disadvantaged users (cf. "affirmative actions") is a more complex and difficult topic and often remains unaddressed. For digital humanism, this presents an opportunity to engage in a multidisciplinary discourse on the design of future recommender systems.

Although the objective of a fairness-focused system is commonly labeled as "fairness," it is crucial to recognize that achieving universal fairness is unattainable. Fairness is a multifaceted issue that is subject to social debates and disagreements, making it impossible to fully resolve. The existence of competing notions of fairness, the diverse requirements of multiple stakeholders, and the fact that fairness is inherently subjective and debatable are reasons for it (Ekstrand et al., 2022).

Emerging approaches aim to address fairness-related issues in recommender systems (Boratto & Marras, 2021). As already sketched before in recommender systems, there are various stakeholders with different fairness concerns (Ekstrand et al., 2022):

- *Consumer fairness* involves treating users fairly and ensuring no systematic disadvantages.
- *Provider fairness* focuses on treating content creators fairly, giving them equal opportunity for their work to be recommended.
- *Subject fairness* is the fair treatment of the people or entities that the recommended items are about.

While fairness concerns for these stakeholders are typically considered separately, some work aims to analyze or provide fairness for multiple stakeholders simultaneously. To promote fairness in recommender systems, it is crucial to identify and address specific harms, understand the stakeholders involved, and contribute to building systems that promote equity and avoid discrimination (Ekstrand et al., 2022). Ideally, responsibility for these tasks is taken by the platforms providing the recommendation services. An overview of different works approaching fairness metrics in ranking and recommendation tasks is given by Patro et al. (2022).

Other venues to design for fairness could be found in a better understanding of users, their values, and motivations. Hence, future studies could delve into psychological theories and empirical studies to understand individuals' preferences and their association with contextual information, personality, and demographic characteristics. Recommender systems are designed to assist human decision-making. Additionally, group recommender systems leverage social psychology constructs to provide recommendations beneficial for groups. While current recommender systems provide useful recommendations, they often lack interpretability and fail to incorporate the underlying cognitive reasons for user behavior (Wilson et al., 2020). This is discussed in the next section on relevant and promising research directions.

6 Human- and Value-Centered Recommender Systems

6.1 Psychology-Informed Recommender Systems

For instance, a survey on psychology-informed recommender systems by Lex et al. (2021) identifies three categories in which different streams from psychological research are being integrated: cognition-inspired, personality-aware, and affect-aware recommender systems. Cognition-inspired recommender systems employ models from cognitive psychology to enhance the design and functionality of recommender systems. Personality-aware recommender systems consider individual personality traits to alleviate cold-start situations for new users and improve

personalization by increasing recommendation list diversity. For instance, the widely used Five-Factor Model (FFM), also known as the Big Five model or the OCEAN model, is often applied in recommender systems research to describe human personality traits (McCrae & John, 1992). Neidhardt et al. (2015) introduced a picture-based approach to elicit user preferences in tourism. Tourism products are complex, and users often have difficulty expressing their needs, especially in the early stages of the travel decision process. The approach introduces seven factors that combine the FFM and tourism roles from literature and creates a mapping between the factors and pictures. In this way, pictures can be used to implicitly and nonverbally elicit the preferences of users and allow users to interact with the RS in a more enjoyable way. Additionally, affect-aware recommender systems consider the emotional state and affective responses of users to provide more tailored recommendations.

With these approaches aiming to better describe the user, one needs to remain aware that these ultimately highly indirect methods to derive human traits and emotions are often built upon strongly debated theories in psychology (see also below) and that their validity is very limited due to technological shortcomings, assumptions, and negligence (Agüera y Arcas et al., 2017). Whether these research directions therefore actually constitute progress in building more "human-centered" systems or are yet another unsuitable attempt that effectively "dehumanizes" users and violates their privacy needs to be painstakingly observed and investigated. As such, from a digital humanist's perspective, this research direction of recommender systems needs to be met with caution.

6.2 Value-Oriented Recommender Systems

Lately, researchers have been attempting to create more moral and human-centered recommender systems that are in line with human values and support human welfare. In order to create recommender systems that reflect human values and advance well-being, Stray et al. (2022) advocate incorporating human values into them. They stress the importance of taking an interdisciplinary approach to this task. The psychological mechanisms that drive changes in user behavior, including their needs and individual abilities to cope with uncertainty and ambiguous situations, are frequently overlooked (FeldmanHall & Shenhav, 2019).

However, it is important to acknowledge that the field of recommender systems research tends to overly rely on easily accessible and quantifiable data, often neglecting discussions on the stability of observable attitudes and behaviors over time ("McNamara Fallacy"; Jannach & Bauer, 2020) and the potential for interventions to bring about change. Many of the prevailing psychological theories and concepts in the quickly developing field of recommender systems are based on early psychological research (such as Ekman's theory of basic emotion, 1992), which has since frequently been shown to be oversimplified and unable to adequately capture the complex and dynamic nature of human attitudes, behaviors,

cognition, and emotion (Barrett, 2022). To illustrate, the stability of a person's personality across different situations has been challenged, as individuals do not consistently behave in accordance with their inner urges (Montag & Elhai, 2019). Montag and Elhai also highlight that while longitudinal studies have demonstrated the overall stability of personality over time, subtle changes can occur, and life events impact personality development. This knowledge emphasizes the importance of considering the context in psychology-aware recommender systems. Integrating these insights into recommender systems could provide a more nuanced understanding of users' preferences and behaviors.

6.3 Embodiment in Recommender Systems

Additionally, recent advancements in cognitive science shine a light on the intricate relationship between the decision-making processes and brain-body functions, which holds significance for the design and functionality of recommender systems. Renowned psychologist and cognitive scientist Lisa Feldman Barrett emphasizes the brain's role in maintaining the body's vital resources, referred to as allostasis, to facilitate various cognitive and physical activities (Barrett, 2017). Considering these insights, it becomes evident that incorporating an understanding of brain-body functions is crucial in the design of recommender systems. Acknowledging the interplay between cognitive processes and physiological regulation allows for a more holistic approach to recommendation algorithms. However, it is essential to recognize the characteristics of human decision-making, its potential as well as its vulnerabilities (Turkle, 2022). To illustrate, while a serendipitous recommendation might fit the user's profile perfectly, their emotional state might simply not allow them to receive it as such (Nguyen et al., 2018). Furthermore, some users' personalities are more, others less accepting of serendipitous recommendations.

In summary, recent discoveries in cognitive science, including the understanding of brain-body functions and decision-making processes, have direct implications for the design and improvement of recommender systems. Integrating insights from cognitive psychology and neuroscience can enhance the accuracy and relevance of recommendations.

6.4 Trust in Recommender Systems

The interaction between a user and a recommender system is also defined by the amount of trust the user holds against it. The more a user trusts the recommender system to generate useful items, the more the user will accept the items (Harman et al., 2014). This is especially important when recommending serendipitous items, as these may appear unexpected which can lead to trust issues (Afridi, 2019). Providing a user with relevant recommendations will establish trust over time, while providing unsatisfying recommendations will erode trust. There are also

other challenges; according to Ricci et al. (2022, p. 7) "some users do not trust recommender systems, thus they play with them to see how good they are at making recommendations," and they noted that "a certain system may also offer specific functions to let the users test its behavior in addition to those just required for obtaining recommendations."

Recent evidence highlighting the importance of autonomy support for human well-being and positive outcomes has raised concerns regarding autonomy within technology design (Calvo et al., 2020). However, incorporating design strategies that promote human autonomy faces two major challenges. Firstly, the breadth of designing for autonomy is extensive, as technologies now play a role in various aspects of our lives, such as education, workplace, health, and relationships, spanning different stages of human development. Secondly, these design practices present ethical dilemmas that challenge existing conceptions of autonomy across disciplines, particularly considering that most technologies are designed to influence human behaviors and decision-making processes.

6.5 Socially Responsible Designs

The inclusion of "socially responsible designs" (Heitz et al., 2022, p. 2) in research and development programs could open opportunities to create recommenders that result in actions and choices that are advantageous to both individuals and society (e.g., Stray et al., 2021). Incorporating individual-level elements and user characteristics, psychology-aware recommender systems can provide a fresh viewpoint in the field of recommender systems research. These systems aim to offer more individualized, varied, and interpretable recommendations by making use of psychological categories and ideas. To ensure a more thorough understanding of user preferences and behavior in the design of recommender systems, additional exploration and consideration of the trait-state perspectives and research development in psychology and the cognitive sciences of human characteristics, intervention possibilities, and the impact of social context are required.

7 Conclusions

We have provided but a glimpse into the area of recommender systems, their importance for the modern Web, and their potential impact on individuals and democracy. Following an overview of techniques used in recommender systems and strategies to evaluate and optimize them, we have focused on the ongoing research discussions dealing with the topics of diversity and fairness. From these discussions, the following take-away messages emerge:

- Optimizing systems for historic patterns and behavior data can indicate effectiveness and improvements of systems that in fact lead to decreasing user satisfaction and narrowing of utility. Other aspects, such as diversity in the results, even if

they are not considered correct according to the chosen accuracy-oriented evaluation measures, are important to judge the quality of a system.
- When deployed in areas relevant to democracy, such as news and media, or for the well-being and success of individuals, such as job recommendation, values defined by society shall be given preference over the objectives of service providers, for instance, by means of policy and regulation. Operationalizing these values is challenging but imperative.
- Recommendation settings are complex tasks involving multiple stakeholders. Questions such as diversity and fairness must always be addressed from their diverse and mostly conflicting points of view. Again, whose interests are to be prioritized should ultimately be decided by society or the affected community. Interdisciplinary approaches are required to define concepts such as fairness, e.g., by involving political scientists and others. These are challenging and complex tasks that ultimately require approaches that model societal values. Currently, these are open issues despite the growing body of work addressing these topics.
- Not every research direction dealing with human features is human centered. In fact, there is a chance that they are not even scientific as they are often built on very weak assumptions, spurious effects, and insufficient technology. Conclusions drawn based on such systems are not only invalid but potentially harmful as they can build the basis for decisions that affect individuals. As such, poorly designed and careless research poses the risk of building "de-humanizing" systems, rather than providing the claimed "human centricity."

For digital humanism, recommender systems are a central technology. They are information filters and automatic decision systems. They make content accessible and at the same time act as opaque gatekeepers. They serve humans as well as business interests. They can be shaped according to values—including those of Digital Humanism.

Discussion Questions for Students and Their Teachers

1. Select an area where recommender systems are used and identify stakeholders.
   - For each stakeholder, discuss how they would benefit from a concept of diversity if applied to the recommender system.
   - Which concept would that be?
   - How would this connect to a notion of fairness from their perspective?
   - Which values could they follow and how would that affect their goals and the definitions chosen?
   - Where do the interests of different stakeholders align?
2. Recommender systems are necessary to efficiently navigate the vast amounts of online data and content; at the same time they are a normative factor and can be used to exert control and power. Discuss the usefulness and threats imposed by recommender systems. Collect anecdotal evidence of success stories of recommenders, failures, and concerns and identify individually desired functions of improved, future recommenders and platforms.
3. For technical solutions, a model of the real world and the operationalization of functions and goals is necessary. Discuss how human and societal values could be modeled and operationalized to enable more fair systems.

Learning Resources for Students

For a deeper understanding of the inner workings and principles of recommender systems, it is strongly suggested to directly refer to the Recommender Systems Handbook (3rd edition), in particular the chapters on techniques, applications, and challenges; novelty and diversity; multistakeholder systems; and fairness in recommender systems:

1. Ricci, F., Rokach, L., & Shapira, B. (2022). Recommender systems: Techniques, applications, and challenges. In Recommender Systems Handbook, 3rd ed., 1–35. DOI: 10.1007/978-1-0716-2197-4_1.
2. Castells, P., Hurley, N., & Vargas, S. (2022). Novelty and diversity in recommender systems. In Recommender Systems Handbook, 3rd ed., 603–646. DOI: 10.1007/978-1-0716-2197-4_16.
3. Abdollahpouri, H. & Burke, R. (2022) Multistakeholder Recommender Systems. In Recommender Systems Handbook, 3rd ed., 647–677. DOI: 10.1007/978-1-0716-2197-4_17.
4. Ekstrand, M. D., Das, A., Burke, R., & Diaz, F. (2022). Fairness in recommender systems. In Recommender Systems Handbook, 3rd ed., 679–707. DOI: 10.1007/978-1-0716-2197-4_18.

Critical takes on current practices and methodology in recommender systems and machine learning research can be found in:

1. Jannach, D., & Bauer, C. (2020). Escaping the McNamara fallacy: towards more impactful recommender systems research. AI Magazine, 41(4):79–95.
2. Agüera y Arcas, B., Mitchell, M., & Todorov, A. (2017). Physiognomy's New Clothes. Medium. URL: https://medium.com/@blaisea/physiognomys-new-clothes-f2d4b59fdd6a

For a broader, multi-perspective discussion on the topics of diversity, fairness, and value-based recommendation, the following articles will provide additional input:

1. Helberger, N., Karppinen, K., & D'Acunto, L. (2018). Exposure diversity as a design principle for recommender systems, Information, Communication & Society, 21(2):191–207. DOI: 10.1080/1369118X.2016.1271900.
2. Binns, R. (2018) Fairness in Machine Learning: Lessons from Political Philosophy. Conference on Fairness, Accountability, and Transparency. Proceedings of Machine Learning Research, 81:1–11.
3. Stray, J., Halevy, A., Assar, P., Hadfield-Menell, D., Boutilier, C., Ashar, A., Beattie, L., Ekstrand, M., Leibowicz, C., Sehat, C. M., Johansen, S., Kerlin, L., Vickrey, D., Singh, S., Vrijenhoek, S., Zhang, A., Andrus, M., Helberger, N., Proutskova, P., Mitra, T., & Vasan, N. (2022). Building Human Values into Recommender Systems: An Interdisciplinary Synthesis. arXiv preprint arXiv:2207.10192.

Acknowledgments This work was supported by the Christian Doppler Research Association (CDG). This research was funded in whole, or in part, by the Austrian Science Fund (FWF) [P33526]. For the purpose of open access, the authors have applied a CC BY public copyright license to any author accepted manuscript version arising from this submission.

References

Abdollahpouri, H., & Burke, R. (2022). Multistakeholder recommender systems. In *Recommender systems handbook* (3rd ed., pp. 647–677). Springer. https://doi.org/10.1007/978-1-0716-2197-4_17

Afridi, A. H. (2019). Transparency for beyond-accuracy experiences: A novel user interface for recommender systems. *Procedia Computer Science, 151*, 335–344. https://doi.org/10.1016/j.procs.2019.04.047

Agüera y Arcas, B., Mitchell, M., & Todorov, A. (2017). *Physiognomy's new clothes. Medium*. https://medium.com/@blaisea/physiognomys-new-clothes-f2d4b59fdd6a

Aridor, G., Goncalves, D., & Sikdar, S. (2020). Deconstructing the filter bubble: User decision-making and recommender systems. In *Proceedings of the 14th ACM conference on recommender systems (RecSys '20)* (pp. 82–91). ACM. https://doi.org/10.1145/3383313.3412246.

Barrett, L. F. (2017). The theory of constructed emotion: An active inference account of interoception and categorization. *Social Cognitive and Affective Neuroscience, 12*(1), 1–23. https://doi.org/10.1093/scan/nsw154

Barrett, L. F. (2022). Context reconsidered: Complex signal ensembles, relational meaning, and population thinking in psychological science. *American Psychologist, 77*(8), 894–920. https://doi.org/10.1037/amp0001054

Binns, R. (2018). Fairness in machine learning: Lessons from political philosophy. Conference on Fairness, Accountability, and Transparency. *Proc Mach Learn Res, 81*, 1–11. https://doi.org/10.48550/arXiv.1712.03586

Björneborn, L. (2017). Three key affordances for serendipity: Toward a framework connecting environmental and personal factors in serendipitous encounters. *Journal of Documentation, 73*(5), 1053–1081. https://doi.org/10.1108/JD-07-2016-0097

Boratto, L., & Marras, M. (2021). Advances in bias-aware recommendation on the web. In *Proceedings of the 14th ACM international conference on web search and data mining* (pp. 1147–1149). https://doi.org/10.1145/3437963.3441665

Burke, R. (2007). Hybrid web recommender systems. *The adaptive web: Methods and strategies of web personalization* (pp. 377–408). https://doi.org/10.1007/978-3-540-72079-9_12

Calvo, R. A., Peters, D., Vold, K., & Ryan, R. M. (2020). Supporting human autonomy in AI systems: A framework for ethical enquiry. *Ethics of digital well-being: A multidisciplinary approach* (pp. 31–54). https://doi.org/10.1007/978-3-030-50585-1_2

Castells, P., Hurley, N., & Vargas, S. (2022). Novelty and diversity in recommender systems. In *Recommender systems handbook* (3rd ed., pp. 603–646). https://doi.org/10.1007/978-1-0716-2197-4_16.

Chouldechova, A., & Roth, A. (2020). A snapshot of the frontiers of fairness in machine learning. *Communications of the ACM, 63*(5), 82–89. https://doi.org/10.1145/3376898

Dahlgren, P. (2021). A critical review of filter bubbles and a comparison with selective exposure. *Nordicom Review, 42*, 15–33. https://doi.org/10.2478/nor-2021-0002

Ekman, P. (1992). An argument for basic emotions. *Cognition and Emotion, 6*(3–4), 169–200. https://doi.org/10.1080/02699939208411068

Ekstrand, M. D., Das, A., Burke, R., & Diaz, F. (2022). Fairness in recommender systems. In *Recommender systems handbook* (3rd ed., pp. 679–707). https://doi.org/10.1007/978-1-0716-2197-4_18.

FeldmanHall, O., & Shenhav, A. (2019). Resolving uncertainty in a social world. *Nature Human Behaviour, 3*(5), 426–435. https://doi.org/10.1038/s41562-019-0590-x

Harman, J. L., O'Donovan, J., Abdelzaher, T., & Gonzalez, C. (2014). *Dynamics of human trust in recommender systems* (pp. 305–308). In Proceedings of the 8th ACM Conference on Recommender systems (RecSys '14). https://doi.org/10.1145/2645710.2645761

Heitz, L., Lischka, J. A., Birrer, A., Paudel, B., Tolmeijer, S., Laugwitz, L., & Bernstein, A. (2022). Benefits of diverse news recommendations for democracy: A user study. *Digital Journalism, 10*(10), 1710–1730. https://doi.org/10.1080/21670811.2021.2021804

Helberger, N., Karppinen, K., & D'Acunto, L. (2018). Exposure diversity as a design principle for recommender systems. *Information, Communication and Society, 21*(2), 191–207. https://doi.org/10.1080/1369118X.2016.1271900

Helberger, N. (2019). On the democratic role of news recommenders. *Digital Journalism, 7*(8), 993–1012. https://doi.org/10.1080/21670811.2019.1623700

Jannach, D., & Bauer, C. (2020). Escaping the McNamara fallacy: Towards more impactful recommender systems research. *AI Magazine, 41*(4), 79–95. https://doi.org/10.1609/aimag.v41i4.5312

Jannach, D., Zanker, M., Felfernig, A., & Friedrich, G. (2010). *Recommender systems: An introduction*. Cambridge University Press.

Jannach, D., Resnick, P., Tuzhilin, A., & Zanker, M. (2016). Recommender systems—Beyond matrix completion. *Communications of the ACM, 59*(11), 94–102. https://doi.org/10.1145/2891406

Kaminskas, M., & Bridge, D. (2016). Diversity, serendipity, novelty, and coverage: A survey and empirical analysis of beyond-accuracy objectives in recommender systems. *ACM Transactions on Interactive Intelligent Systems, 7*(1), 1–42. https://doi.org/10.1145/2926720

Keijzer, M. A., & Mäs, M. (2022). The complex link between filter bubbles and opinion polarization. *Data Science, 5*(2), 139–166. https://doi.org/10.3233/DS-220054

Lee, S. Y., & Lee, S. W. (2022). Normative or effective? The role of news diversity and trust in news recommendation services. *International Journal of Human–Computer Interaction, 39*(6), 1216–1229. https://doi.org/10.1080/10447318.2022.2057116

Lex, E., Kowald, D., Seitlinger, P., Tran, T. N. T., Felfernig, A., & Schedl, M. (2021). Psychology-informed recommender systems. Foundations and Trends®. *Information Retrieval, 15*(2), 134–242. https://doi.org/10.1561/1500000090

Liu, P., Zhang, L., & Gulla, J. A. (2023). *Pre-train, prompt and recommendation: A comprehensive survey of language modelling paradigm adaptations in recommender systems*. arXiv preprint https://doi.org/10.48550/arXiv.2302.03735.

McCrae, R. R., & John, O. P. (1992). An introduction to the five-factor model and its applications. *Journal of Personality, 60*(2), 175–215. https://doi.org/10.1111/j.1467-6494.1992.tb00970.x

Michiels, L., Leysen, J., Smets, A., & Goethals, B. (2022). What are filter bubbles really? A review of the conceptual and empirical work. In *Adjunct proceedings of the 30th ACM conference on user modeling, adaptation and personalization* (pp. 274–279). https://doi.org/10.1145/3511047.3538028.

Möller, J., Trilling, D., Helberger, N., & van Es, B. (2018). Do not blame it on the algorithm: an empirical assessment of multiple recommender systems and their impact on content diversity. *Information, Communication and Society, 21*(7), 959–977. https://doi.org/10.1080/1369118X.2018.1444076

Montag, C., & Elhai, J. D. (2019). A new agenda for personality psychology in the digital age? *Personality and Individual Differences, 147*, 128–134. https://doi.org/10.1016/j.paid.2019.03.045

Neidhardt, J., Seyfang, L., Schuster, R., & Werthner, H. (2015). A picture-based approach to recommender systems. *Information Technology and Tourism, 15*, 49–69. https://doi.org/10.1007/s40558-014-0017-5

Nguyen, T. T., Hui, P.-M., Harper, F. M., Terveen, L., & Konstan, J. A. (2014). Exploring the filter bubble: The effect of using recommender systems on content diversity. In *Proceedings of the 23rd international conference on World wide web* (pp. 677–686). https://doi.org/10.1145/2566486.2568012.

Nguyen, T. T., Maxwell Harper, F., Terveen, L., & Konstan, J. A. (2018). User personality and user satisfaction with recommender systems. *Information Systems Frontiers, 20*, 1173–1189. https://doi.org/10.1007/s10796-017-9782-y

Pariser, E. (2011). *The filter bubble: What the internet is hiding from you*. Penguin Press.

Patro, G. K., Porcaro, L., Mitchell, L., Zhang, Q., Zehlike, M., & Garg, N. (2022). Fair ranking: A critical review, challenges, and future directions. In *Proceedings of the 2022 ACM conference*

on fairness, accountability, and transparency (pp. 1929–1942). https://doi.org/10.1145/3531146.3533238.

Raza, S., & Ding, C. (2020). A regularized model to trade-off between accuracy and diversity in a news recommender system. In *2020 IEEE international conference on big data* (pp. 551–560). https://doi.org/10.1109/BigData50022.2020.9378340.

Ribeiro, M. T., Ziviani, N., Moura, E. S. D., Hata, I., Lacerda, A., & Veloso, A. (2015). Multiobjective pareto-efficient approaches for recommender systems. *ACM Transactions on Intelligent Systems and Technology, 53*, 1–20. https://doi.org/10.1145/2629350

Ricci, F., Rokach, L., & Shapira, B. (2022). Recommender systems: Techniques, applications, and challenges. In *Recommender systems handbook* (3rd ed, pp. 1–35). https://doi.org/10.1007/978-1-0716-2197-4_1.

Smets, A. (2023). Designing for serendipity: A means or an end? *Journal of Documentation, 79*(3), 589–607. https://doi.org/10.1108/JD-12-2021-0234

Smets, A., Michiels, L., Bogers, T., & Björneborn, L. (2022). Serendipity in recommender systems beyond the algorithm: A feature repository and experimental design. In *Proceedings of the 9th joint workshop on interfaces and human decision making for recommender systems co-located with 16th ACM conference on recommender systems* (pp. 44–66). https://ceur-ws.org/Vol-3222/paper4.pdf

Stray, J., Vendrov, I., Nixon, J., Adler, S., & Hadfield-Menell, D. (2021). What are you optimizing for? Aligning recommender systems with human values. *CoRR,* abs/2107.10939. https://doi.org/10.48550/arXiv.2107.10939

Stray, J., Halevy, A., Assar, P., Hadfield-Menell, D., Boutilier, C., Ashar, A., Beattie, L., Ekstrand, M., Leibowicz, C., Sehat, C. M., Johansen, S., Kerlin, L., Vickrey, D., Singh, S., Vrijenhoek, S., Zhang, A., Andrus, M., Helberger, N., Proutskova, P., Mitra, T., & Vasan, N. (2022). *Building human values into recommender systems: An interdisciplinary synthesis*. arXiv preprint https://doi.org/10.48550/arXiv.2207.10192.

Turkle, S. (2022). *The empathy diaries: A memoir*. Penguin.

Vermeulen, J. (2022). To nudge or not to nudge: News recommendation as a tool to achieve online media pluralism. *Digital Journalism, 10*, 1–20. https://doi.org/10.1080/21670811.2022.2026796

Werthner, H., et al. (2019). *The Vienna manifesto on digital humanism*. https://dighum.org/dighum-manifesto/

Werthner, H., Stanger, A., Schiaffonati, V., Knees, P., Hardman, L., & Ghezzi, C. (2023). Digital humanism: The time is now. *Computer, 56*(1), 138–142. https://doi.org/10.1109/MC.2022.3219528

Wilson, J. R., Gilpin, L., & Rabkina, I. (2020). *A knowledge driven approach to adaptive assistance using preference reasoning and explanation*. arXiv preprint https://doi.org/10.48550/arXiv.2012.02904.

Zhang, S., Yao, L., Sun, A., & Tay, Y. (2019). Deep learning based recommender system: A survey and new perspectives. *ACM Computing Surveys, 52*(1), 1–38. https://doi.org/10.1145/3285029

SPEAKER

Second African Digital Humanism Summer School, Accra, Ghana, 31 August – 05 September 2026

Carlo Ghezzi

Carlo Ghezzi is an Emeritus Professor at Politecnico di Milano, where he has been teaching and doing research for over 40 years. He is an ACM Fellow, IEEE Fellow, member of Academia Europaea, and member of the Italian Academy of Sciences (Istituto Lombardo). He received the ACM SIGSOFT Outstanding Research Award. He has been President of Informatics Europe. He has done research on programming languages and software engineering. He has published over 200 papers in international journals and conferences and co-authored 6 books. He is interested in the ethical implications of research in computer science. He is currently chairing the Ethics Committee at Politecnico di Milano.

Lecture 7a: *Software Engineering After AI*

Reader article:

Carlo Ghezzi, Masoud Ebrahimi, Damir Isovic, and Marjan Sirjani: Breaking Disciplinary Silos: The Case of Software Engineering.
In L. Hagedorn, U Schmid, S. Winter, S. Woltran (Eds.): Proceedings First Interdisciplinary Science and Research Conference DIGHUM 2025, LNCS 16319, pp. 421-435. Springer Nature, Cham Switzerland, 2026. https://doi.org/10.1007/978-3-032-11108-1_32

Amel Bennaceur, Carlo Ghezzi, Jeff Kramer, and Bashar Nuseibeh: Responsible Software Engineering – Requirements and Goals.
In H. Werthner, C. Ghezzi, J. Kramer, J. Nida-Rümelin, B. Nuseibeh, E. Prem, A. Stanger (Eds.): Introduction to Digital Humanism – A Textbook, pp. 299-315. Springer Nature, Cham, Switzerland, 2024. https://doi.org/10.1007/978-3-031-45304-5_20

Breaking Disciplinary Silos: The Case of Software Engineering

Carlo Ghezzi[1(✉)], Masoud Ebrahimi[2(✉)], Damir Isovic[2], and Marjan Sirjani[2]

[1] DEIB, Politecnico di Milano, Milano, Italy
carlo.ghezzi@polimi.it
[2] Mälardalen University, Västerås, Sweden
masoud.ebrahimi@mdu.se

Abstract. Digital Humanism aims to influence the complex interplay between technology and humankind, striving for a better society while fully respecting universal human rights. While software is the crucial technological component in this interplay, current mainstream approaches in software engineering research and education are largely agnostic about the goals of Digital Humanism. Scientists and engineers tend to view, design, and study software systems from a purely technical perspective, often disregarding human and societal implications. In this position paper, we argue that research and education should be profoundly revisited by broadening the disciplinary landscape through cross-disciplinary collaborations, particularly with the humanities and social sciences, to ensure that the human and societal implications of new designs, as well as the principles of Digital Humanism, are properly considered. We provide an initial roadmap for this effort, mainly addressing software engineering. The roadmap is intended as a basis for further discussion and refinement, and serves as a call for increased community action in this area.

Keywords: Digital Humanism · Human & Societally Critical Systems · Software Engineering Research · Education · Artificial Intelligence · Awareness · Responsible Design · Transparency · Multidisciplinarity

1 Introduction

Software is the defining technology of the 21st century, transforming how humans live and operate. We now inhabit a cyber-physical world, seamlessly integrated and orchestrated by software. Software defines, embeds, and enforces the mechanisms and laws that govern our world: how humans interact with physical entities, how they engage and cooperate with each other, and how society is organized at all levels (e.g., political, industrial, educational, social, and more). Without exaggeration, most software developed and deployed worldwide today is *human and societally critical (HSC)*. This paper focuses on HSC software.

L. Hagedorn et al. (Eds.): DIGHUM 2025, LNCS 16319, pp. 421–435, 2026.
https://doi.org/10.1007/978-3-032-11108-1_32

The societal nature of most software applications has long been under scrutiny by social and political scientists. In the late 1990s, Lawrence Lessig wrote a landmark book, in which he observed that software (code, in his words) is the regulator of the new world [20].
Quoting from Lessig:

> *This code, or architecture, sets the terms on which life in cyberspace is experienced. It determines how easy it is to protect privacy, or how easy it is to censor speech. It determines whether access to information is general or whether information is zoned. It affects who sees what, or what is monitored. In a host of ways that one cannot begin to see unless one begins to understand the nature of this code, the code of cyberspace regulates.*

Lessig indirectly highlights the responsibilities that software engineers have toward humans and society, as their design decisions are embedded in software. Recent advances in AI further amplify the regulatory role of software in HSC systems and raise crucial concerns about societal implications. AI enables the development of larger and more complex software systems, with agents that can behave autonomously, learn and evolve without human intervention, and even be generated automatically. As a result, systems may exhibit unanticipated behaviors. Complete control over such systems is unattainable, and the illusion of control that software developers often have can be dangerous. Lessig's early observations, which Software Engineering (SE) research and education have largely overlooked, can no longer be ignored.

To address these issues, we argue that SE research and education must be grounded in the principles of **Digital Humanism**[1] (DigHum), placing human values, rights, and dignity at the center of technological development. DigHum asserts that technology must serve human and society (not vice versa) by integrating ethical, social, cultural, and philosophical considerations into the design and governance of digital systems. For further discussion, see [17–19,30–32].

We need to revisit how software for HSC systems is designed[2]: How are goals related to human and societal values identified and set? How are they explicitly and precisely (ideally, even formally) specified? How are they engineered into software? How can they be inspected? How can we ensure they hold, even as the system evolves? Meeting these challenges requires traditional SE technical knowledge to be combined with multidisciplinary expertise. Likewise, academic education should teach how to center human and societal values in software design and engage students in practical, multidisciplinary settings.

The challenges of DigHum for HSC systems can only be met by breaking down the disciplinary silos that have compartmentalized SE research and edu-

[1] Vienna Manifesto on Digital Humanism https://caiml.org/dighum/dighum-manifesto/, published in May 2019, defines DigHum and justifies the need for it. Its publication inspired initiatives and further elaborations of DigHum principles and practices.

[2] In SE, the term "design" typically refers to a specific phase of development. Here, it denotes the entire engineering process: conception, development, deployment, and evolution of a software system.

cation, hindering collaboration with other fields, primarily the humanities and social sciences, but also domain-specific areas like health sciences, political science, and law. Interdisciplinary design deliberations are essential to ensure that relevant human and societal aspects are properly considered and that DigHum principles are upheld. Rethinking education also requires methods to identify common misconceptions and biases among students, and to propose new pedagogical strategies, including real-world scenarios and inclusive design practices, to foster responsible, transparent, and socially beneficial software development.

Interdisciplinarity is often claimed but hard to achieve. It continues to face challenges within academia; e.g., misaligned objectives, divergent terminologies, and methodological incompatibilities. To systematically address these obstacles, a shared framework of values and priorities, such as that proposed by DigHum, is needed. By breaking down disciplinary silos, we can enable future generations of scientists and practitioners to communicate and collaborate more effectively. This is essential for addressing the complexity and challenges of the digital world.

In this paper, we promote *responsible design* as the cornerstone of SE for DigHum and argue that this can only be achieved by including interdisciplinary collaboration in software development from the outset, and by educating future engineers in interdisciplinary teamwork. Designers must take responsibility for their choices and understand the consequences of each decision for humans, society, and the environment. The level of system autonomy and the degree of freedom granted to users must be made explicit and clearly communicated. The pros and cons of each design decision should be critically evaluated, available options transparently presented, and associated responsibilities clearly assigned. This brings clarity of *intentions* and *transparency* to the design process. These concepts also apply to users in a broad sense: it must be clear which user actions lead to which outcomes, ensuring that responsibility and accountability are well understood on both sides.

The paper is structured as follows. Section 2 elaborates on the notion of responsible design and the role of multidisciplinary contributions. Section 3 outlines the main research areas that need to be explored to re-found design methods in the light of DigHum goals. Section 4 discusses the implications of DigHum for rethinking SE education. Section 5 highlights the main contributions of other research groups who are also rethinking software research and education from a humanistic and social perspective. Finally, Sect. 6 offers conclusions and calls on other disciplines to break their own silos. Digitalization is transforming all fields, and students in every discipline need to learn how to communicate with engineers just as much as engineers need to learn to communicate with them. More broadly, the general public must also increase its awareness and responsibility regarding the digital world in which we all live.

2 Software Engineering, Responsible Design, and Ethics

Software engineers are passionate about technology and its potential to shape the world. Their skills have led to innovations that have revolutionized society. SE research has provided engineers with a wide range of conceptual and

practical tools—theories, processes, methods, notations, languages, components, and frameworks—to support the production and evolution of quality software. Despite software's broad and deep impact on humans and society, research has mostly treated software as a purely technical artifact, focusing on traditional objectives applied to any industrial-quality product. The qualities considered are mainly technical, such as functional correctness and performance (speed and efficient use of resources), as well as economic factors, like productivity, portability, or reuse. Human-related qualities are mainly considered under "usability" and viwed as part of human-machine interaction.

Technical knowledge is necessary but not sufficient for developing DigHum-compliant HSC systems. This section examines the foundational principles that should guide in re-founding design of HSC systems from a DigHum perspective. Our discussion is informed by, and closely aligned with, the contributions presented in [31], particularly those by [2,6,17,22,24,26,27,34].

Additional to technical knowledge, *awareness* of the broader context in which the HSC system is embedded is essential for design. Awareness enables designers to state their *intent* explicitly, allowing them to take clear *responsibility* for their work. Such awareness is crucial to ensure that the envisioned HSC system complies with DigHum principles. To achieve this awareness, it is imperative to inform and drive the design process with diverse viewpoints, needs, interests, and concerns that illuminate the broader context in which the artifact will be embedded. This necessitates integrating *multidisciplinary knowledge and argumentation* into the design process. For example, designing an HSC system for assisted living (e.g., serving elderly or impaired individuals) requires input from medical professionals, caregivers, psychologists, sociologists, and ethicists. These contributions provide essential human and societal domain knowledge and clarify the required ethical to ensure DigHum compliance. All participants in the design process must share collective responsibility for the system and its potential actions.

The notion of *intent* is key. Unlike physical entities found in nature, artifacts result from deliberate, creative human activity. Every creative act has an intent, which may be either *conscious* or *unconscious*. An unconscious intent becomes conscious when it is made explicit and precisely formulated. A crucial part of design is making intents explicit by answering questions such as: Why is the new HSC system being designed? What do we aim to achieve? What are the *intended* (i.e., desirable) outcomes? Can other *unintended* effects also occur, and how can we counter (or mitigate) them? By making intents explicit and precise, we can counter the *opacity* that (regrettably) characterizes most software designs. Instead, design *transparency*[3] must be achieved. By this we mean that:

- the design provides a rigorous description of the artifact's intents;
- the artifact is open for inspection by designers and all relevant stakeholders;

[3] The term is the opposite of opacity. It is also closely related to *self-descriptiveness*, defined by the IEEE Standard Glossary of SE Terminology [15] as the degree to which a system or component contains enough information to explain its objectives and properties. This, in turn, relates to *maintainability*, *testability*, and *usability*.

– evidence is provided that the artifact complies with the stated intents and that proper measures are in place to counter undesirable unintended intents.

To make intents explicit, HSC designers inevitably need to confront ethical concerns. *Ethics* is the branch of philosophy concerned with systematic reflection on what is moral by studying what is morally good or bad, and right or wrong. These distinctions are rarely binary; they require a deliberative process to determine the appropriate kind of morality. During design, ethical questions guide the exploration of the human and societal context of the system. Intents must be examined not only for their technical innovation, relevance, and feasibility, but most importantly for how they affect the people involved and their relationships, and what is ultimately good or right (or bad or wrong) for them.

The principles of awareness and responsibility apply not only to designers, but also to all other relevant stakeholders who are involved with or affected by the artifact in use: from decision makers to end-users. Decision makers must take responsibility for adopting the HSC system in society and may need to regulate its use. End-users must provide informed consent for their use. Supporting awareness and responsibility for all stakeholders is a primary design goal; how to achieve this is itself an intent that must be explicitly and precisely stated. We discuss next how design should be rethought to address these issues.

3 Revisiting Design of HSC Systems

This section outlines a research agenda to align responsible software design with DigHum goals. We provide an initial breakdown of areas to be investigated. This list is not exhaustive and should be seen as a starting point for further discussion.

3.1 Design Process

The design process must host the multidisciplinary knowledge that inspires and drives the development of HSC systems. To make intents explicit (i.e., clearly connected to the human and societal context and to ethical implications of the intended consequences) and to ensure compliance with DigHum principles, it is essential to involve all necessary knowledge sources from the start. The process should then be structured to support effective and efficient interdisciplinary collaboration throughout the design activity.

To further motivate the need for multidisciplinary design, consider the implications of the following DigHum principle for HSC systems:

The system shall be fair and behave in an unbiased way.

A biased system is one that shows a benevolent inclination toward (or prejudice against) certain persons, groups, or interests in a way that is considered unfair and thus morally unacceptable. The design process should guide software development to guarantee fairness. Achieving fairness in software systems requires ethical decisions to counter possible manifestations of undesired bias

within complex environments. Sometimes, bias may be benevolent: it is ethically acceptable and may become a deliberate design choice. For example, a system may be biased toward groups that need protection, such as offering special facilities to support minorities or people with impairments. More often, bias is morally unacceptable and may result from a variety of deliberate or unanticipated design choices. Bias can stem from unrepresentative data, objective functions misaligned with societal values, flawed algorithms, or lack of context in modeling. Recent cases have shown that unacceptable bias can arise from ML components trained on biased data [23]. More generally, ML components are intrinsically biased by the data used for their training.

The design process must be structured so that deliberations about fairness inform all design decisions. Input from disciplines beyond SE is essential to identify unacceptable bias and discuss countermeasures, especially from those representing groups potentially affected by unfair system behavior. To address these issues, it is important to precisely, and even formally, define bias in a way that allows for its evaluation. We need methods and tools to identify and reduce undesired bias, and to redefine model goals with community values. Iterative and incremental development processes, involving diverse stakeholders with diverse competences and providing diverse viewpoints, are likely to be the most suitable solutions to accommodate the needed multidisciplinary argumentation and ethical deliberations, as indicated by existing work in this field, presented in Sect. 5.

Software is a living, continuously evolving artifact. Design should account for the future evolution of HSC software. Mechanisms that can detect vulnerabilities to existing values, adapt to changes, and implement countermeasures may enable systems to evolve and adapt over time, even autonomously.

3.2 From Intents to Requirements Specifications and Verification

Intents must ultimately be expressed as requirements that a system must satisfy. Requirements engineering, a subfield of SE, focuses on requirements elicitation and specification. Existing methods are largely agnostic about DigHum-inspired goals[4]. More research is needed to develop a conceptual model and precise, possibly formal, definitions for *DigHum-inspired requirements*, including their dependencies and relationship with technical requirements.

SE research has long focused on formal specifications that are amenable to mechanical manipulation, including automated verification and other assurance techniques. More research is needed to apply these approaches to DigHum-inspired requirements. The earlier example of fairness and bias highlights the

[4] Research papers on requirements appear regularly in the main SE research literature, such as journals (e.g., IEEE Transactions on SE and the ACM Transactions on SE and Methodology) and conferences (e.g., International Conference on SE and the ACM International Conference on the Foundations of SE). There are also specialized venues focusing on requirements engineering (e.g., Springer Nature Requirements Eng. Journal and the IEEE International Requirements Eng. conference).

need for formalization. Among the analyses supported, it is also important to assess whether unintended and undesirable situations can occur. Designers from disciplines beyond SE can make valuable contributions here. The ultimate goal is to achieve *anticipatory design*: anticipating potential problems or threats and identifying countermeasures before issues arise. Anticipatory design aims to reduce reliance on patches that address threats only after they have already caused harm. Systems should also be designed to facilitate or automate the discovery of potential problems during operation and to provide automated corrective adaptation or prompts for designers to make necessary changes.

A key challenge in designing HSC systems is translating high-level human values (e.g., autonomy, dignity, and accountability) into actionable and verifiable software requirements. In the following, we illustrate this with a brief case study.

Free Will. The notion of free will is a complex philosophical concept that raises questions about human agency, autonomy, and the ethical boundaries of machine intervention in human decision-making. In the context of DigHum, free will is not just a philosophical abstraction but a practical concern that must be addressed in the design process.

Preserving human free will and self-determination is a core DigHum principle. Technologies, especially intelligent or autonomous systems, must serve human interests, not the other way around. This means humans should remain the primary decision-makers, empowered rather than constrained by digital tools. However, operationalizing a concept as philosophically complex as free will is challenging, particularly when determining the appropriate boundaries for machine intervention. In many real-world scenarios, it is unclear when a system should defer to human choices and when it must act to prevent harm.

Defining the boundary between human and machine intervention highlights why multidisciplinary perspectives must inform the design process. Engineers often favor full and increased autonomy, which can have valid justifications. However, social scientists may caution that excessive automation can demotivate, deskill, or demote humans to roles they find unacceptable. Conversely, engineers may argue that the simplistic principle that "humans must always retain ultimate control over autonomous system actions" can be just as problematic as relying entirely on system autonomy. For example, in a safety-critical flight control system, overriding the automatic system by a human can led to catastrophic consequences, as shown by the Germanwings crash caused by a suicidal co-pilot.

One possible resolution is to shift the focus from trying to define free will in absolute terms to specifying concrete conditions under which system intervention is ethically permissible. This can be achieved by introducing requirements grounded in the concepts of *malicious intent* and *negligence*. By integrating these notions early in the SE process, we can define explicit constraints that govern system behavior. For example, the system could be required to minimize interference with human decisions, unless it detects malicious intent from the human actor or determines that

failing to intervene would result in a safety-critical outcome due to negligence. This requirement-driven approach reframes the problem: instead of defining autonomy in absolute terms, it identifies ethically and socially grounded exceptions, enabling the design of systems that respect human agency while safeguarding against harm.

To enforce such a framework, the system must include mechanisms that inform users of its operational state and the potential consequences of their actions. Responsibility is grounded in transparency: if a human actor is warned and fully understands the implications of proceeding, yet continues to act in a way that knowingly causes harm, this behavior constitutes malicious intent. If the actor ignores warnings or fails to take remedial action in response to foreseeable dangers, the situation may be considered negligence. In both cases, the system should provide clear and traceable records of its communications and the user's responses, and take preventive measures to avoid safety-critical outcomes. When all available courses of action lead to a safety-critical result, responsibility must be traced through formal traceability and verification. These mechanisms allow developers, regulators, and stakeholders to reconstruct the chain of design decisions, user interactions, and system responses, ensuring accountability is fairly attributed and informing future improvements in ethical system design.

3.3 Improving Design Transparency

Design transparency is essential for achieving collective responsibility among designers of HSC systems. It enables inspection and awareness of all aspects for which designers are accountable. Transparency has long been a major concern for large and complex software systems. Nearly 40 years ago, Turing Award winner Frederick Brooks wrote an influential paper [7] warning about the "essential complexity" of large software systems, which remains a fundamental obstacle to full transparency. Today's software, often incorporating AI-generated components, makes the situation even more challenging. Combating opacity is one of SE's most fundamental challenges.

The ability to specify intents precisely, and to scrutinize and assess them through multidisciplinary argumentation, is key to achieving transparency during design. Transparency is also crucial for other stakeholders (e.g., policy makers, regulators, and end-users), who must take responsibility for the HSC system according to their specific roles. Transparency means enabling stakeholders to ask questions about what the artifact does and how it operates, and to receive precise answers that are both comprehensible and guaranteed to be true.

The traditional SE notions of *modelling, specification, and verification* need to be broadened to achieve transparency through *assured guarantees*. A variety of approaches, notations, and languages have been developed to precisely describe systems: their structure, their behavior, and their properties. Likewise, methods have been devised to verify (or prove) whether an artifact satisfies these descriptions. More research is needed to expand the range of modelling and specification approaches for HSC systems to include DigHum-specific requirements, new ways

to express them precisely or formally, and new ways to provide assurances at all levels, including design, evolution, and use, and for all possible stakeholders.

3.4 The Role of AI in HSC Systems

AI plays an increasingly crucial role in HSC systems design, from code generation based on natural language prompts, to learned components produced by training models, to autonomous agents and assistants used during design or embedded within HSC systems. The SE research has already been transformed by advances in AI, as seen in the rapid growth of AI-related papers in the SE literature in recent years. However, it is important that this research always keeps DigHum principles as primary concerns.

For example, let us return to the crucial notions of transparency. AI can threaten transparency by making parts of HSC systems difficult to understand, predict, guarantee, or explain[5]. We need to determine where, how, and under which circumstances this lack of transparency can be accepted, and how it affects the guarantees that can be assured. More broadly, we must develop an understanding of when the automation and autonomy introduced by AI is beneficial. This involves addressing the ethical question of what constitutes the right or good kind of AI, and what constitutes the wrong or bad kind of AI [1,11].

4 Implications on Education

DigHum requires transformative educational practices to prepare students to engage meaningfully with HSC systems. Central to this is the ability to understand the implications of technology for humans and society. These insights should be systematically incorporated into updated curricula that bring ethics to developers, engineers, and data scientists, making it clear that becoming a responsible professional is a recognized goal of their education.

A bottom-line requirement is that all students in technological fields, especially software engineering, should take a basic course on ethics as a necessary part of their education, just like mathematics or physics. In addition, technical courses should systematically highlight situations where design decisions can have ethical implications. Another key skill for students is the ability to collaborate across disciplinary boundaries. Traditional degree programs often isolate students within their own domain, but complex societal challenges require education that fosters integration between technology, social sciences, and humanities from the beginning. Therefore, it is essential to give students opportunities to learn how to work with peers from different backgrounds, including social sciences, humanities, and ethics.

A successful example of effective multidisciplinary collaboration in education is the Solar Car project at our university. In this project, 17 students from various disciplines collaborated to build a solar-powered vehicle that later competed in the Solar Challenge in Australia. By working together on a common

[5] Explainable AI is widely recognized as a fundamental research challenge.

goal over a one-year period, the team gradually developed a shared vocabulary and methodology. This process was further supported by a compulsory 10-week course for all students, which covered theoretical frameworks such as agile software development and intercultural communication, and included guest lectures and mentoring from industry professionals.

The Solar Car project demonstrated how students from various backgrounds can co-create solutions that are both technically and socially aware. However, such efforts should not remain isolated. Multidisciplinary perspectives need to be integrated into teaching and learning in a more systematic and natural way, reflecting the real challenges of the digital society. The next step is to formalize a methodology for multidisciplinary software education that incorporates ethics and human factors directly into the curriculum. Such a framework would enable the systematic inclusion of DigHum in technical education.

4.1 Operationalizing Ethics and Values in Education

To support these educational goals, ethical frameworks can provide foundational structure. The *ACM Code of Ethics* [3] and the *IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems* [16] establish principles such as fairness, accountability, and transparency, which can be embedded in project work and evaluation criteria. In addition, Value-Based Engineering (VBE) offers a concrete methodology for translating abstract values into design requirements [25]. In an educational context, VBE can serve as a tool for students to practice aligning system design with stakeholder values and social goals, making ethics a part of technical reasoning rather than an afterthought.

Integrating ethical and value-driven approaches into coursework can help foster a mindset of responsible innovation. Assignments and group projects should include ethical impact assessments and value articulation exercises, encouraging students to consider trade-offs and societal implications early in their careers.

4.2 Toward Generalizable Educational Frameworks

The Solar Car example demonstrates how education can be reimagined to support DigHum. To generalize its success, we need to identify repeatable patterns and embed them into curricula and teaching materials. This includes designing interdisciplinary capstone projects, evaluating student work through ethical perspectives, and creating shared resources and infrastructure across institutions. Scalability across cultural and institutional contexts can be achieved through modular teaching units, collaborative platforms, and cross-institutional faculty training. In addition, documenting and sharing effective practices will help motivate others to make necessary changes in their curricula. Rather than mandating change, it is better to recognize and reward teachers who integrate DigHum into their courses. This will require updated mechanisms for evaluating academics that foster interdisciplinarity.

To conclude, higher education institutions should take a leading role in preparing future generations of professionals. These professionals should understand their roles and responsibilities in shaping the direction of the digital world they help create, and should align their work with the principles of DigHum.

5 Related Work

Autonomous decision-making in HSC systems requires alignment with Social, Legal, Ethical, Empathetic, and Cultural (SLEEC) values [5,8]. De Sanctis et al. [8] argue that contemporary software quality models must be redefined to incorporate SLEEC concerns as first-class quality attributes, rather than treating them as external considerations. This rethinking is further motivated by recent regulatory developments, such as the EU AI Act, which mandates demonstrable adherence to ethical principles. However, as noted by Morley et al. [21], in the absence of formal specification mechanisms, these mandates risk remaining aspirational, subject to selective interpretation and inconsistent implementation, which can undermine both trust and compliance.

To mitigate such risks, formal methods for capturing and enforcing ethical considerations have gained traction. Townsend et al. [28] propose a methodology that enables the translation of pluralistic normative principles, expressed in natural language, into actionable SLEEC-aligned operational rules. Other efforts explore the use of formal logics, particularly deontic logic, to encode normative constructs such as obligations, prohibitions, and permissions. For example, Arkoudas et al. [4] demonstrate how mechanized deontic logic can enable robots to verify ethical rules and generate justifiable decisions, laying foundational work for ethical reasoning in autonomous agents.

Dennis et al. [9] extend a Beliefs-Desires-Intentions (BDI) framework, implemented using the Gwendolen language [10], for agent-based systems by incorporating ethical constraints. They define a logic-based ethical operator that governs decision-making and resolves moral dilemmas by prioritizing ethical rules according to a predefined hierarchy. Their case studies show how logic-based verification can be applied to real-world HSC contexts. However, it remains an open problem whether the ethical rules were objectively formalized, or the hierarchies used to resolve moral dilemmas actually adhere to regulatory and ethical norms.

To address these issues, ontology-based requirements engineering has been used to study, model, and specify ethical concepts and constraints. For example, Guizzardi et al. [13] and Houghtaling et al. [14] propose approaches for structuring ethical requirements using ontologies. While these methods offer semantic clarity and traceability, they differ in their level of abstraction and currently lack standardized formalization languages for automated verification.

Despite these advancements, the field still lacks a unified, formalized framework that defines and operationalizes trust principles within SE. Autili et al. [5] address this gap by proposing a research roadmap aimed at aligning SE practice with SLEEC-aware societal responsibilities. To structure this vision, they introduce four archetypal human roles in relation to digital systems: proactive

agents, reactive responders, passive experiencers, and trust participants. These roles help frame the diverse and often conflicting demands placed on autonomous systems. Building on this conceptual foundation, the roadmap defines four primary research domains: software development processes, requirements engineering, software architecture and design, and verification and validation. Each domain emphasizes the integration of responsible-by-design principles, ensuring that HSC and SLEEC considerations are embedded from the earliest stages of system conception. This line of work closely aligns with what we propose here.

The notion of a *software exoskeleton* has recently been introduced to serve as a personalized ethical interface, mediating the user's digital experience according to their declared values [29]. Building on this idea, De Sanctis and colleagues proposed a reference architecture for ethically aware autonomous systems, formalizing the roles of human actors, agents, and governance mechanisms. Their methodology includes a logical framework for compiling and reasoning over SLEEC rules [29], which helps enable transparent and inspectable decision-making processes. Together, these works provide a comprehensive and technically grounded foundation for advancing ethical compliance in HSC systems.

Another related and relevant line of research, investigated at the Bavarian Institute for Digital Transformation [12,33,34], focuses on how ethical deliberations can become effective drivers of the software development process. The referenced papers report positive experiences with adopting incremental and iterative Agile processes, which foster collaboration and adaptability. Some of these approaches show that integrating ethical considerations into established development frameworks is both feasible and beneficial, supporting the alignment of SE practices with societal values and regulatory requirements.

6 Conclusions

Software has revolutionized the way we live, work, and interact, becoming an integral part of our daily lives. However, the rapid advancement of technology has outpaced our understanding of its societal implications. As we continue to develop and deploy software systems, especially those that are human and societally critical, it is important to re-evaluate our approach to SE research, education, and practice. This paper provides an initial reflection on how to align research and education in SE with the goals and principles of DigHum, articulated through the following actions:

- *Promote DigHum principles:* Encourage and support interdisciplinary and multidisciplinary collaborations amongst computer scientists, ethicists, sociologists, and experts from other relevant disciplines to drive responsible design of DigHum-compliant HSC systems.
- *Develop a DigHum-inspired SE research agenda:* Develop new process models that effectively integrate multidisciplinary contributions in the design of HSC software, with the goal of making intents explicit and fostering responsible design. Research is also needed to improve transparency at all levels, supporting rigorous specifications and guaranteed assurances of DigHum-inspired requirements throughout design and for all stakeholders.

- *Integrate DigHum into SE education:* Curricula must include interdisciplinary courses covering ethics, social sciences, and humanities to prepare future engineers for the societal impacts of their work. Students must also learn how to practice interdisciplinary collaboration through innovative project courses.
- *Engage with all stakeholders, including policy makers and the public:* Foster dialogue between technologists, policy makers, and the public to ensure that software systems serve the broader interests of society.

By embracing these actions, we can guide the field of SE toward a future that not only advances technology but also upholds the values and rights of humanity.

Acknowledgments. This work would not exist without the pioneering and inspiring collective efforts of all contributors to the Digital Humanism initiative. We are indebted to them for laying the foundations for this discussion in the Vienna Manifesto on Digital Humanism and in the subsequent activities. Part of the work is supported by the ASSURE project, part of the multi-disciplinary research initiative at MDU.

References

1. Acemoglu, D., Restrepo, P.: The wrong kind of AI? artificial intelligence and the future of labour demand. Cambridge J. Regions, Econom. Soc. **13**(1), 25–35 (2019). https://doi.org/10.1093/cjres/rsz022
2. Akkermans, H.: The social responsibilities of scientists and technologists in the digital age, pp. 65–81. Springer Nature Switzerland, Cham (2024). https://doi.org/10.1007/978-3-031-45304-5_5
3. Anderson, R.E.: ACM code of ethics and professional conduct. Commun. ACM **35**(5), 94–99 (1992)
4. Arkoudas, K., Bringsjord, S., Bello, P.: Toward ethical robots via mechanized deontic logic. In: AAAI Fall Symposium on Machine Ethics, pp. 17–23 (2005)
5. Autili, M., De Sanctis, M., Inverardi, P., Pelliccione, P.: Engineering digital systems for humanity: a research roadmap. CoRR abs/2412.19668 (2024). https://doi.org/10.48550/ARXIV.2412.19668
6. Bennaceur, A., Ghezzi, C., Kramer, J., Nuseibeh, B.: Responsible software engineering: requirements and goals, pp. 299–315. Springer Nature Switzerland, Cham (2024). https://doi.org/10.1007/978-3-031-45304-5_20
7. Brooks, F.P.: No silver bullet essence and accidents of software engineering. Computer **20**(4), 10–19 (1987). https://doi.org/10.1109/MC.1987.1663532
8. De Sanctis, M., Inverardi, P., Pelliccione, P.: Do modern systems require new quality dimensions? In: Bertolino, A., Faria, J.P., Lago, P., Semini, L. (eds.) QUATIC'24. CCIS, vol. 2178, pp. 83–90. Springer (2024). https://doi.org/10.1007/978-3-031-70245-7_6,
9. Dennis, L., Fisher, M., Slavkovik, M., Webster, M.: Formal verification of ethical choices in autonomous systems. Robot. and Auton. Syst. **77**, 1–14 (2016). https://doi.org/10.1016/j.robot.2015.11.012, https://www.sciencedirect.com/science/article/pii/S0921889015003000
10. Dennis, L.A., Farwer, B.: Gwendolen: A BDI language for verifiable agents. In: AISB'08, pp. 16–23 (2008)

11. Dubhashi, D.: Can universities combat the 'wrong kind of AI'? Commun. ACM **65**(12), 24–26 (2022). https://doi.org/10.1145/3522710
12. Gogoll, J., Zuber, N., Kacianka, S., Greger, T., Pretschner, A., Nida-Rümelin, J.: Ethics in the software development process: from codes of conduct to ethical deliberation. Philosophy Technol. **34**(4), 1085–1108 (2021)
13. Guizzardi, R., Amaral, G.C.M., Guizzardi, G., Mylopoulos, J.: An ontology-based approach to engineering ethicality requirements. Softw. Syst. Model. **22**(6), 1897–1923 (2023). https://doi.org/10.1007/S10270-023-01115-3
14. Houghtaling, M.A., et al.: Standardizing an ontology for ethically aligned robotic and autonomous systems. IEEE Trans. Syst. Man Cybern. Syst. **54**(3), 1791–1804 (2024). https://doi.org/10.1109/TSMC.2023.3330981
15. IEEE: IEEE standard glossary of software engineering terminology (1990). https://doi.org/10.1109/IEEESTD.1990.101064
16. IEEE Standards association and others: the IEEE global initiative on ethics of autonomous and intelligent systems (2018). https://1standards.ieee.org/developlindconn/ec/autonomous_systems.html
17. Lee, E.A.: Are we in control?, pp. 165–174. Springer Nature Switzerland, Cham (2024). https://doi.org/10.1007/978-3-031-45304-5_11
18. Lee, E.A.: Certainty vs. intelligence. In: Bridging the Gap Between AI and Reality. AISoLA 2024. vol. LNCS 15217 (2024). https://doi.org/10.1007/978-3-031-75434-0_2
19. Lee, E.A.: The coevolution: the entwined futures of humans and machines. The MIT Press (2020). https://doi.org/10.7551/mitpress/12307.001.0001
20. Lessig, L.: Code and Other Laws of Cyberspace. Basic Books Inc, USA (1999)
21. Morley, J., Floridi, L., Kinsey, L., Elhalal, A.: From what to how: an initial review of publicly available ai ethics tools, methods and research to translate principles into practices. Sci. Eng. Ethics **26**(4), 2141–2168 (2019). https://doi.org/10.1007/s11948-019-00165-5
22. Nida-Rümelin, J., Staudacher, K.: Philosophical foundations of digital humanism, pp. 17–30. Springer Nature Switzerland, Cham (2024). https://doi.org/10.1007/978-3-031-45304-5_2
23. Rudin, C., Wang, C., Coker, B.: The Age of secrecy and unfairness in recidivism prediction. Harvard Data Sci. Rev. **2**(1) (2020). https://hdsr.mitpress.mit.edu/pub/7z10o269
24. Schiaffonati, V.: Promises and perils in moralizing technologies, pp. 255–265. Springer Nature Switzerland, Cham (2024). https://doi.org/10.1007/978-3-031-45304-5_17
25. Spiekermann, S., Winkler, T.: Value-based engineering with IEEE 7000. IEEE Technol. Soc. Mag. **41**(3), 71–80 (2022)
26. Sun, F., Isovic, D.: Responsible AI under the philosophical framework of Digial Humanism. Int. J. Inf. Theor. Appl. **32**(4) (2025). http://www.es.mdu.se/publications/7286-
27. Tamburrini, G.: Artificial intelligence and large-scale threats to humanity, pp. 241–254. Springer Nature Switzerland, Cham (2024). https://doi.org/10.1007/978-3-031-45304-5_16
28. Townsend, B.A., et al.: From pluralistic normative principles to autonomous-agent rules. Minds Mach. **32**(4), 683–715 (2022). https://doi.org/10.1007/S11023-022-09614-W
29. Troquard, N., De Sanctis, M., Inverardi, P., Pelliccione, P., Scoccia, G.L.: Social, legal, ethical, empathetic, and cultural rules: Compilation and reasoning. In:

Wooldridge, M.J., Dy, J.G., Natarajan, S. (eds.) AAAI'24, pp. 22385–22392 (2024). https://doi.org/10.1609/AAAI.V38I20.30245
30. Werthner, H.: The vienna manifesto on digital humanism. In: Digital transformation and ethics, pp. 338–357. Ecowin (2020)
31. Werthner, H., et al.: Introduction to digital humanism. Springer Nature Switzerland (2024)
32. Werthner, H., Prem, E., Lee, E.A., Ghezzi, C.: Perspectives on digital humanisms. International series of monographs on physics, Springer Chams (2021). https://doi.org/10.1007/978-3-030-86144-56
33. Zuber, N., Gogoll, J., Kacianka, S., Pretschner, A., Nida-Rümelin, J.: Empowered and embedded: ethics and agile processes. Human. Soc. Sci. Commun. **9**(1), 1–13 (2022)
34. Zuber, N., Gogoll, J., Kacianka, S., Nida-Rümelin, J., Pretschner, A., et al.: Value-sensitive software design: ethical deliberation in agile development processes. Hannes Werthner Carlo Ghezzi Jeff Kramer Julian Nida-Rümelin Bashar Nuseibeh Erich Prem, p. 339 (2024)

Responsible Software Engineering: Requirements and Goals

Amel Bennaceur, Carlo Ghezzi, Jeff Kramer, and Bashar Nuseibeh

Abstract In this chapter, we provide an introduction to the discipline of requirements engineering as part of the software engineering process. We indicate how to elicit, articulate, and organize the goals of complex software systems as an explicit expression of the requirements that the proposed or existing software system is expected to achieve and maintain, including what the system should avoid performing. We advocate that system requirements goals can and should be used to explicitly capture, express, and reason about the diverse digital humanism values which are of concern in socio-technical systems. This is an essential aspect of responsible software engineering.

1 Introduction

Software is creating a new digital world in which humans live, individually and socially. This is a large and complex *socio-technical system* where the boundaries between digital, physical, and social spaces are increasingly disappearing. Many activities in such a system are automated, supporting and sometimes replacing human work and creating new functionalities that did not exist before. Humans interact with software-enabled agents in their daily life. Software now defines and administers most of the laws that govern the world. This was observed in the late 1990s by Lawrence Lessig, in his framing of "Code is Law" (Lessig, 2000).

A. Bennaceur (✉) · B. Nuseibeh
The Open University, Milton Keynes, UK

Lero, University of Limerick, Limerick, Ireland
e-mail: amel.bennaceur@open.ac.uk; bashar.nuseibeh@open.ac.uk

C. Ghezzi
DEIB, Politecnico di Milano, Milan, Italy
e-mail: carlo.ghezzi@polimi.it

J. Kramer
Imperial College London, London, UK
e-mail: j.kramer@imperial.ac.uk

H. Werthner et al. (eds.), *Introduction to Digital Humanism*,
https://doi.org/10.1007/978-3-031-45304-5_20

Software engineers, who create code, are the *demiurges*. Although they are responsible for "technical" decisions, the consequences of their decisions go far beyond the purely technical sphere, often with unintended and unanticipated consequences. At the same time, legal systems have lagged behind in adapting to technological changes.

How can the implications of socio-technical systems developed by software engineers be properly considered when systems are conceived and developed? How can the values and issues of digital humanism drive the software engineering process? How does that engineering process interact with other processes (political, normative, etc.) both ex ante—while the system is designed—and ex post, when systems are deployed and operate?

Engineers have traditionally focused on functional correctness, efficiency, and scalability of their solutions. By and large, they have ignored fairness, inclusivity, and deep consideration of the social implications of their solutions. They have mastered technology and make complex technical decisions, but rarely consider the consequences of future use and misuse of their products in society.

In this chapter, we advocate that these issues and values must be considered and explicitly integrated into the software engineering process: in particular, in the explicit expression of the requirements that the proposed or existing system is expected to achieve and maintain. This focus on so-called requirements engineering can provide a bridge between the world in which digital humanism values arise and the digital machine that software engineers design, build, and deploy in that world.

We begin with an overview of requirements engineering, focusing on ways in which goals and requirements are elicited from diverse stakeholders and how they can be explicitly modeled and analyzed. We illustrate how such goal models can be extended to capture various human values and discuss how they can be analyzed for the purpose of validation and verification. We conclude with a discussion on a more responsible software engineering discipline and some suggested exercises to engage students in the articulation of and reflection on digital humanism goals in software systems.

2 Requirements Engineering (RE)

RE has been the subject of several popular books and surveys; this section gives a brief introduction to requirements as a primary basis for sound software engineering. It also provides relevant references for further exploration of the area.

2.1 Introduction to RE

> *Requirements engineering is the branch of software engineering concerned with the real-world goals for, functions of, and constraints on software systems. It is also concerned with the relationship of these factors to precise specifications of software behavior, and to their evolution over time and across software families.* (Zave, 1997)

This definition by Zave emphasizes that a new software system is introduced to solve a real-world problem and that a good understanding of the problem and the associated context is at the heart of RE. Therefore, it is important not only to define the *goals* of the software system but also to *specify its behavior* and to understand the *constraints and the environment* in which this software system will operate. The definition also highlights the need to consider change, which is inherent in any real-world situation. Finally, the definition suggests that RE aims to capture and distill the experience of software development across a wide range of applications and projects.

Although Zave's definition identifies some of the key challenges in RE, the nature of RE itself has been changing. First, RE is not specific to software alone but to *socio-technical systems* in general, of which software is only a part. Software today permeates every aspect of our lives, and therefore, one must not only consider the technical but also the physical, economical, and social aspects. Second, an important concept in RE is *stakeholders*, that is, individuals or organizations who stand to gain or lose from the success or failure of the system to be constructed (Nuseibeh & Easterbrook, 2000). Stakeholders play an important role in eliciting requirements as well as in validating them.

While the definition of the requirements helps delimit the solution space, the requirement problem space is less constrained, making it difficult to define the environment boundary, negotiate the resolution of conflicts, and set acceptance criteria (Cheng & Atlee, 2007). Therefore, several guidelines are given to define and regulate the RE processes in order to build adequate requirements (Robertson & Robertson, 2012). Figure 1 summarizes the main activities of RE:

Elicitation. Requirements elicitation aims to discover the needs of stakeholders as well as understand the context in which the system-to-be will operate. It may also explore alternative ways in which the new system could be specified. Several techniques can be used including (i) traditional data gathering techniques (e.g., interviews, questionnaires, surveys, analysis of existing documentation), (ii) collaborative techniques (e.g., brainstorming, workshops, prototyping), (iii) cognitive techniques (e.g., protocol analysis, card sorting), (iv) contextual techniques (e.g., ethnographic techniques, discourse analysis), and (v) creativity techniques (e.g., creativity workshops, facilitated analogical reasoning).

Modeling. The results of the elicitation activity often need to be described precisely and in a way accessible by domain experts, developers, and other stakeholders. A wide range of techniques and notations can be used to represent requirements, ranging from informal to semi-formal to formal (mathematical) methods. The choice of the appropriate method often depends on the kind of analysis or reasoning that needs to be performed.

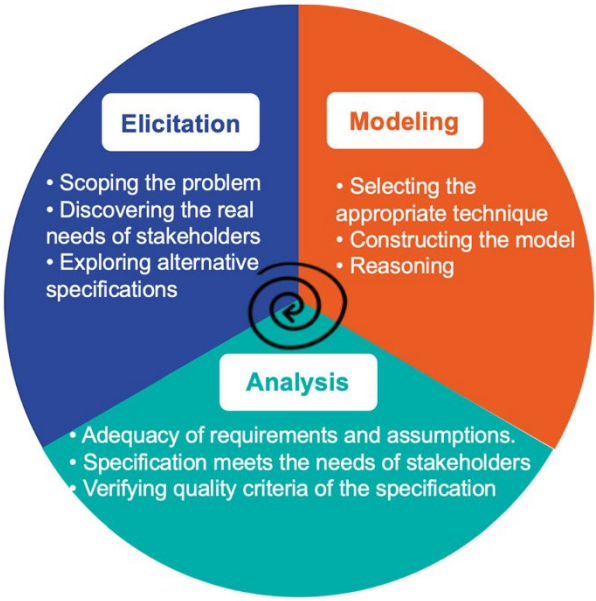


Fig. 1 Main activities of requirements engineering

Analysis and Assurance. Requirements quality assurance seeks to identify, report, analyze, and fix defects in requirements. It involves both validation and verification. Validation aims to check the adequacy of the modeled requirements and domain assumptions with the actual expectations of stakeholders. Verification covers a wide range of checks including quality criteria of the modeled requirements (e.g., consistency).

2.2 Requirements and Goals

Zave and Jackson (1997) suggest that there are three main kinds of artifacts that requirements engineers would produce during the RE activities:

- Statements about the *domain*, describing properties that are true regardless of the presence or actions of the machine (or software system)
- Statements about *requirements*, describing properties that the stakeholders want to be true of the world in the presence of the machine
- Statements about the *specification*, describing what the machine needs to do to achieve the requirements

These statements can be written in natural language, formal logic, semi-formal languages, or indeed some combination of them, and Zave and Jackson are not prescriptive about that. What is important is their relationship: *The specification of the machine, together with the properties of the domain, should satisfy the requirements.*

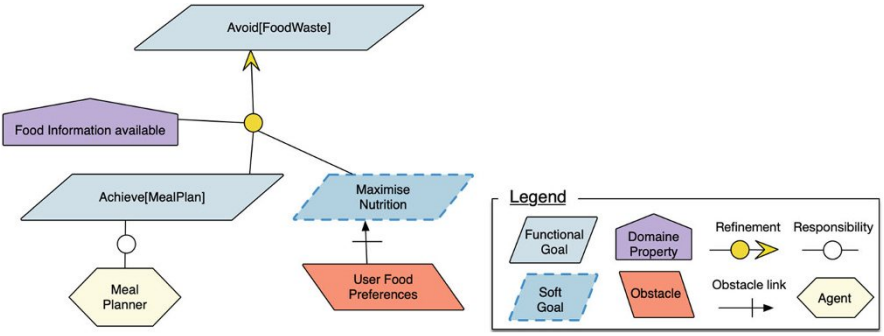


Fig. 2 A sample goal model for the MealPlanning example

To illustrate those notions, let us consider an example of *meal planning* as a way to tackle an important and pressing societal problem, food waste. It is estimated that food waste per capita by consumers in Europe and North America is 95–115 kg/year. Food waste is often caused by insufficient planning of purchases and consumption by individuals. Effective strategies to reduce wasteful behavior should require minimum time and cognitive effort from consumers. The Feed me, Feed me exemplar (Bennaceur et al., 2016) describes a system based on the Internet of Things to support the production, distribution, and consumption of food. We use ideas and challenges from the Feed me, Feed me exemplar to focus on how our approach can support individuals in reducing food waste in households.

For example, to avoid food waste, we should plan meals. This can be achieved by refinements (as illustrated in Fig. 2). The *Avoid[FoodWaste]* goal is refined into sub-goals and associated domain properties. A *goal* in this model is defined as a prescriptive statement that the system should satisfy through the cooperation of *agents* such as humans, devices, and software. Goals may refer to services to be provided (functional goals) or quality of service (soft goals).

Achieve[MealPlan] is a functional goal, while *MaximiseNutrition* is a soft goal. While functional goals can be satisfied or not, soft goals are often optimized. Keywords such as *Achieve*, *Maintain*, and *Avoid* are used to characterize the intended behaviors of the goals and can guide their formal specification.

Domain properties are descriptive statements about the environment. For example, *Food Information Available* is a domain property. An important relationship is that the goal *Avoid[FoodWaste]* can be satisfied through *Achieve[MealPlan]* and *MaximiseNutrition* assuming *Food Information Available.*

Besides describing the contribution of sub-goals (and associated domain properties) to the satisfaction of a goal, refinement links are also used for the operationalization of goals and assigning them to (software) agents. For example, *MealPlanner* is responsible for satisfying the goal *Achieve[MealPlan].*

Finally, *Conflict* links are used to represent obstacles to the satisfaction of goals. For example, *UserFoodPreferences* may hinder the satisfaction of *MaximiseNutrition*. Reasoning about obstacles enables risk analysis of the goals by eliciting properties that may obstruct the satisfaction of goals.

Hence, RE is grounded in the real world; it involves understanding the *environment (domain)* in which the system-to-be will operate and defining a detailed, consistent *specification* of the software system-to-be. This process is incremental and iterative as illustrated in Fig. 2.

Zave and Jackson specify five clear criteria for this process to complete:

- Each goal has been validated with the stakeholders.
- Each domain property has also been validated with the stakeholders.
- The goal does not constrain the environment or refer to the future.
- There exists a proof of the satisfaction of goals.
- The goals and domain properties are consistent.

2.3 The Need for Human-Centered Values

The essence of RE is a good understanding of problems, which includes analyzing the domain, communicating with stakeholders, and preparing for system evolution. However, techniques such as machine learning, automated compositions and interactions, and creativity disrupt the traditional models of software development and call for quicker, if not immediate, response from requirements engineering. Moreover, the social underpinning and the increasing reliance on software systems for every aspect of our life call for better methods to understand the impact and implications of software solutions on individuals and society as a whole.

For example, several pressing global problems such as climate change and sustainability engineering as well as increasingly important domains such as user-centered computing and other inter- and cross-disciplinary problems challenge existing processes and techniques. It is no longer enough to understand the needs of stakeholders and the constraints of the environments in which a software system is deployed; we also need to understand the values of the stakeholders and understand the broader impact of deploying software solutions. In the next section, we move to values and their interaction with requirements.

3 Values We Live By: Eliciting, Articulating, and Organizing Goals

Digital humanism argues for adopting a broader framework where, besides the technical perspective, multiple perspectives (including ethical, social, legal, political, and economic) are considered when developing systems that have an impact on individuals and society.

Recent work has promoted the need to consider ethics and values during the development of software systems (Whittle, 2019). As outlined by Mougouei et al.

(2018), "people are demanding that practitioners align technologies with human values." Some approaches have been proposed to assess and study values in software engineering (Winter et al., 2019), to incorporate social values in software design patterns (Hussain et al., 2018), and to measure the impact of values in requirements engineering activities (Perera et al., 2021). Values are well studied in human-computer interaction and information systems (Cockton, 2004).

For RE, this means rethinking the world in terms of broader and changing stakeholders, their needs, and their values. It also means rethinking the notion of requirements satisfaction to incorporate values and the inevitability of failure and change. Some of the challenges of doing so stems from the subjectivity and uncertainty of values. Values are *subjective* and depend on the diverse viewpoints of stakeholders because different stakeholders describe value requirements differently. As a result, they have different and sometimes contradictory requirements. For example, if we consider the value of fairness, serving a protected group with priority can promote fairness in society, but, at the same time, it may seem discriminatory to others. Values are *uncertain* and are often better understood once the software solution is deployed. For example, awareness of gender bias in data may lead to the deployment of existing equality policies, and their impact and consequences are better understood once deployed.

The debate has long focused on principles and codes of conduct for considering values in software systems. However, it is increasingly moving to tools and processes for implementing those values and principles in practice. While awareness of the potential issues is increasing at a fast rate, the software/requirements engineering community's ability to take action to mitigate the associated risks is still in its infancy. There is still a need to close the gap between principles and practices for engineers to apply ethics at each stage of the development pipeline and to signal to researchers where further work is needed. In other words, we need methods to move from "what" values to embed to "how" those values can be embedded in software systems. This section provides some direction toward achieving this goal.

3.1 Values and RE Activities

Let us first review the RE activities with humanistic values in mind.

Elicitation. Social scientists, ethicists, philosophers, policymakers, technologists, and civil society have been involved in a debate around what is necessary to enable society to capitalize on the opportunities of software systems while ensuring fair and ethical decision-making is maintained. Participatory design aims to elicit the values of multiple stakeholders by following several steps, which include:

- Involving actual users for eliciting value concerns
- Using personas to consider/assume user values
- Using prototypes to analyze assumptions about values
- Using diversity in members selected from various stakeholder groups
- Focusing on cultural sensitivities
- Being considerate of language needs of different stakeholder groups
- Developing empathy with users, emulating their experiences
- Building an atmosphere of trust for stakeholders to voice their opinions
- Applying user feedback to improve mock-ups and prototypes

In addition to continual engagement with stakeholders and practitioners, reflection on practices and the impact of the developing software systems is equally important. The Self-Reflection Tool of the Responsible Research and Innovation (RRI)[1] framework helps practitioners consider the societal and ethical issues that may be involved with technology. Learning by doing underpins the AREA (Anticipate, Reflect, Engage, and Act) approach to RRI. This means that professional and social responsibility is best developed through experience and reflective practice. The guidelines for such practices include:

1. Involving a wide range of actors and people in practice, deliberation, and decision-making. This strengthens democracy and broadens sources of expertise, disciplines, and perspectives.
2. Envisioning impact and reflecting on the underlying assumptions, values, and purposes to better understand how the developed systems shape the future. This yields valuable insights and increases the capacity to act on what we know.
3. Communicating in a meaningful way the methods, results, conclusions, and implications to enable public scrutiny and dialogue. This benefits the visibility and understanding of the developed systems.
4. Being able to modify modes of thought and behavior, overarching organizational structures, in response to changing circumstances, knowledge, and perspectives. This aligns action with the needs expressed by different stakeholders.

Modeling. In Value-Based Requirements Engineering (Thew & Sutcliffe, 2018), values are seen as personal attitudes and beliefs which influence functional and non-functional requirements. There is evidence of human values being treated as software requirements, specifically as soft goals or non-functional requirements (Barn, 2016). In values-first software engineering, Ferrario et al. (2016) argue that complex wicked problems such as sustainability should be treated as soft goals, not as functional requirements. Nurwidyantoro et al. (2022) postulate that non-functional requirements can be seen as a subset of human values and propose to classify human values and align them to system values. They found system value themes, such as efficiency and usability, similar to non-functional requirements.

[1] https://rri-tools.eu/. Accessed 10 April 2023.

Fig. 3 Dimensions to consider for eliciting and operationalizing values

Assurance. Operationalizing values is defined as "the process of identifying human values and translating them to accessible and concrete concepts so that they can be implemented, validated, verified, and measured in software" (Shahin et al., 2022). It is common for stakeholders to gain a better understanding of their values as they experience, reflect, and learn more about them (Gentile, 2010). However, elicitation and modeling approaches focus on early stages of the development process, with little attention given to the satisfaction of values in deployed software systems (Shahin et al., 2022). Software solutions can help stakeholders articulate, measure, and reflect on their values while they are experiencing the software. Values@Runtime (Bennaceur et al., 2023) deal with uncertainty by delaying some decisions until software is in operation. It adopts an adaptive process to engage stakeholders and to support learning about models of stakeholders' values. It provides values instantiation as a means of representing the concrete actions that stakeholders associate with values (Hanel et al., 2017). This framework supports values operationalization in terms of (i) representation, instantiation, and monitoring of values and behavior; (ii) understanding existing mismatches between values and users' behavior based on analysis; and (iii) recommending ways to align values and behavior as well as reflecting on the recommendations.

Hence, eliciting and operationalizing values involves three dimensions (see Fig. 3):

- People, through the adoption of a human-centered view and participatory design as well as involving a diversity of stakeholders and teams
- Artifacts, by making explicit value statements and engineering systems for diverse stakeholders
- Processes, by linking values between requirements and implemented software and by being transparent and open to accountability about implementation practices and mindful of project impact and following current standards and regulations

3.2 Values and Goals

Let us consider the example of fairness when food shopping (Farahani et al., 2021). The high-level goal is *Achieve[FairShopping]* and might be refined in multiple ways—see Fig. 4.

For example, when the domain property *AbundantStock* holds, then the goal is to *maximize Products Sold*, which can be operationalized by allowing users to buy as many products as they want/need. When the domain property Limited Stock holds, then there needs to be a choice between two goals: *Achieve[EquitableAccess]* by prioritizing protected groups or *Achieve[EqualAccess]* by limiting the maximum amount of product per shopper without distinction between shoppers. While not mutually exclusive, the choice is driven by consideration of multiple stakeholders, e.g., supermarkets' willingness to implement different procedures, government's willingness to support protected groups, and public acceptance of prioritizing protected groups. For example, prioritizing a protected group can be perceived as fair for some people, but at the same time, it may seem discriminatory to others. In other words, a goal model can help highlight the stakeholders involved when making value-sensitive choices, e.g., fair for whom or who is responsible for the choice. The goal model helps highlight and contrast alternative operationalization of values.

Emotions can be used as proxy to values and leveraged to design inclusive processes (Hassett et al., 2023). For example, the *Supermarket* might want the stakeholder group, *Vulnerable Shopper* (e.g., older person or person with special needs), to *feel Cared for*, which then leads to prioritizing protected groups.

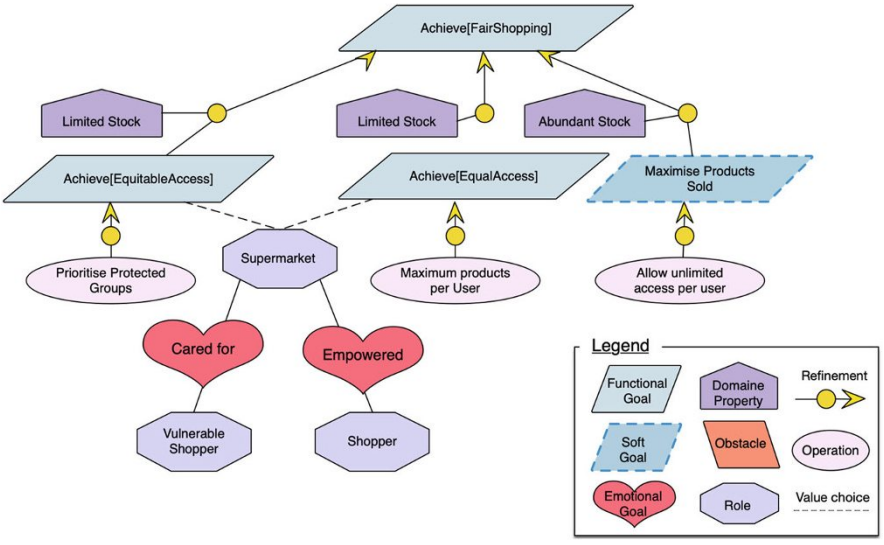


Fig. 4 A sample (emotional) goal model for the fair food shopping example

4 Toward Responsible Software Engineering: DigHum Goals in the Life Cycle

In this section, we discuss how the principles of digital humanism may guide the life cycle of socio-technical systems: from the conception and development of a system, to its operation in the real world, to its continuous evolution.

4.1 Requirements and Other Activities

The first and most important step consists of understanding and articulating the requirements. The previous two sections shed light on this crucial activity, through which developers assume explicit responsibility with respect to the system under development. Through goal models, they express a contract with stakeholders and future users, which states what the system is expected to achieve. Traditionally, software engineers are educated to focus on goals that refer to the functionalities and expected behaviors to be provided by the system and on technical qualities, like efficiency (e.g., average response time of certain transactions), portability of the implementation on different architectures, or security (e.g., guaranteed restricted access to certain data or functionalities). In our context, however, requirements also reflect the general humanistic values, modeled as explicit goals to be met by the future system. For example, fairness is explicitly modeled as a goal to achieve in the context of the food shopping example.

Eliciting and articulating these goals is critical, but also quite difficult and highly context dependent. The technical skills possessed by software engineers alone may fall short. Not only stakeholders and user representatives must be involved, but also experts from other domains—like philosophy (ethics), history, social sciences, economics, or law—may have a lot to say in order to understand goals, analyze and resolve conflicts, and prioritize among them, but also to anticipate possible uses (or misuses) of socio-technical systems in the real world. Depending on the specific system being developed, a deliberative process needs to be put in place, which gives voice to different viewpoints and then responsibly leads to decisions that inform all subsequent development steps.

Requirements are a prerequisite for design and implementation (Fig. 5). These are technical steps that lead to a functioning socio-technical system. Design is responsible for defining the software architecture, i.e., decomposing a system into components and deciding how different components interact and communicate. Implementation is responsible for producing an executable system, often through a combination of programmed parts, libraries, and software frameworks.

However, requirements also permeate many parts of the systems development process. During system design, requirements are used to inform decision-making about different design alternatives. During system implementation, requirements are used to enable system prototyping and testing. Once the system has been deployed,

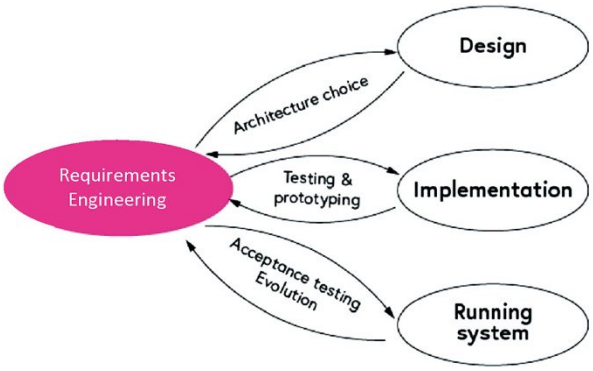


Fig. 5 Crucial relationship between requirements and other activities

requirements are used to drive acceptance tests to check whether the final system does what the stakeholders originally wanted. In addition, requirements are reviewed and updated during the software development process as additional knowledge is acquired and stakeholders' needs are better understood.

Each step of the development process may lead to the definition of additional requirements through a better understanding of the domain and associated constraints. Therefore, there is a need to consider requirements, design, and architecture concurrently, and this is often the process adopted by software engineers.

4.2 Software Processes

Different process models can be followed to guide development, ranging from top-down (waterfall) processes to bottom-up and iterative processes. Waterfall processes are monolithic and sequential: they try to strictly enforce completion of the requirements phase before proceeding to the design phase, which must itself be completed before moving to implementation. Strict sequential ordering of phases is only suitable for highly structured systems that operate in well-defined, formalizable, and highly stable contexts. It is not suitable for ill-defined and unstable settings as is the case in most socio-technical systems, where humans play a fundamental role. More flexible—iterative and incremental—life cycles, such as the popular process models which fall under the term *agile processes*, are almost always adopted for the latter kind of systems. Agile processes, which envision the development of system increments, e.g., via *sprints* in the SCRUM agile methodology (Schwaber & Beedle, 2002), appear as a suitable setting to accommodate the necessary deliberations through which digital humanism-inspired requirements can be explored and then guide development. The chapter by Zuber et al. in this book provides deeper insights into how agile development methods are inherently suitable for embedding digital humanism values into software systems.

4.3 Validation and Verification

Two other important activities need to be carried out during development: verification and validation (V&V). The two terms shed light on two complementary kinds of assurances. *Validation* is the assurance that the system meets the needs of the customer and other identified stakeholders. It often involves acceptance and suitability with external customers. *Verification* is the assurance that the system complies with its specification. The two terms would of course be synonyms if specifications were exhaustive and complete. This is almost inevitably impossible in practice. In addition, as we discuss next, the needs of customers and other stakeholders continuously evolve, and therefore, an upfront complete specification is impossible to realize.

V&V is itself not a stage of system development, but rather a cross-cutting activity that permeates all development steps. Requirements are continuously verified and validated as they are elicited and formalized; likewise, architectures and implementation increments are subject to V&V. Delivery of (partial) applications for real use presupposes an adequate level of V&V to check compliance with specifications, including possible existing regulations, and adherence to users' needs. It is also possible to design systems in a way that these checks are made automatically by the system while it operates, at run-time (run-time V&V).

V&V is practiced through two complementary approaches: systematic reasoning and testing. Systematic reasoning tries to analyze the artifacts under development to prove that the stakeholders' expectations are met by affirming that violations of those expectations are impossible. Testing develops experiments that try to bring the system into desirable and undesirable states, to collect empirical evidence that the system being developed can be delivered for practical use. The two approaches are complementary, since exhaustive testing is impossible to achieve and tools to assist in systematic reasoning do not scale up to large systems.

4.4 The Running System

The life cycle of an application does not end when it is deployed. Most systems, and especially those successfully used in practice, are subject to continuous evolution, traditionally called maintenance. New requirements may arise from real use, pre-existing requirements may need to be adapted due to new insights gained while the system has been in use, opportunities for improvements may be discovered, and errors or other problematic situations that evaded V&V may show up during execution. To support evolution, specifically designed monitors may be implemented in the deployed applications to perform run-time V&V, checking for the insurgence of potential risks, violations of desirable policies, or mishaps.

5 Conclusions

In this chapter, we have explained the central role of requirements engineering in the software engineering process of software production and evolution. We have explained why we advocate requirements specifications and goals as the most promising and pragmatic technique to explicitly express the societal and digital humanism values which are so crucial to sound and responsible software engineering of socio-technical systems. This includes not just what the system should achieve but also what it should avoid performing. Some form of continuous monitoring of the running system will be needed to support assessment as part of responsible software engineering. This will also require that software engineers are involved in outreach activities regarding the global effects of their products: to assess impact, use, and abuse. As mentioned, experts in other disciplines (social scientists, lawyers, etc.) will also need to be involved, not just at the elicitation stage but also when the system is deployed and running. This diversity of stakeholders is becoming more and more important as systems are embedded in society.

We believe that this extension of traditional software engineering to include humanistic values is essential to cope with complex socio-technical systems. It will inevitably require further research, practice, and education to refine the techniques, to gain further empirical evidence and experience, and to ensure dissemination to the profession.

Discussion Questions for Students and Their Teachers
In the following hypothetical projects work together with colleagues from different disciplines and with different backgrounds, to articulate the Digital Humanism goals and overall requirements to be reached by a hypothetical socio-technical system, understanding potential conflicts, and mitigating potential risks, including misuse.

Consideration should also be given to what should be automated and what is left to humans to perform and also whether it is possible to ascertain whether or not the resulting system is compliant with the specified goals and explicit values.

1. Hypothetical Project 1: Citizen Forensics

   The police are overstretched, criminality is on the rise, . . . how can citizens participate in deterring crime and helping the police (and each other) detect anti-social incidents and solve crime.

   Hints/issues: you could explore risk and issues around surveillance (before or after incidents), harassment, privacy, citizen-police relations, and information sharing. . .A resource: https://www.citizenforensics.org
2. Hypothetical Project 2: Technology in the Courtroom

   It's not easy to be a judge. . . It is necessary to assess as correctly as possible whether an offender will recidivate, what sentence is appropriate for the particular offense, whether or not the sentence should be suspended, and much more.

   Wouldn't it be great if technology could make judgments easier?

   Many ethical questions arise here, such as who is "fairer," a technical system or a human? How transparent must a decision by a technical system be that

supports a court ruling? Is such a system more of a science fiction fantasy à la Minority Report or an actual chance to counter prejudices, perception biases, or even racist tendencies among judges?

Here are some papers and articles about this topic:
https://link.springer.com/content/pdf/10.1007/s10506-022-09310-1.pdf
https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm
https://link.springer.com/article/10.1007/s10506-022-09312-z
https://scholarship.law.umn.edu/mlr/58/

Proposed by: Anna Dhungel

3. Hypothetical Project 3: My Truth, Your Truth

Since the corona pandemic, I don't recognize some of my friends. Through some social media forums, they have become vaccination opponents, mask deniers, and world conspirators.... How can social media be made social and responsible again without immediately giving the feeling of living in a "dictatorship of opinion"?

Hint: The following article gives a brief introduction into the democratic roles of news recommender systems: https://doi.org/10.1080/21670811.2019.1623700.

Proposed by: Kian Schmalenbach and Eva Gengler

Learning Resources for Students

1. Van Lamsweerde, A., 2009. Requirements engineering: From system goals to UML models to software (Vol. 10, p. 34). Chichester, UK: John Wiley & Sons.

The book presents a systematic method to elaborate complex system models, analyze them, and derive software specifications from them. The method is known as KAOS (Keep All Objectives Satisfied). The goal models in this chapter used notations and formalisms from this book.

2. Brey, P. and Dainow, B., 2021. Ethics by design and ethics of use in AI and robotics. The SIENNA project-Stakeholder-informed ethics for new technologies with high socioeconomic and human rights impact.

The document provides guidance for including ethical principles and procedures into the design and development processes of AI systems.

3. IEEE Standard Model Process for Addressing Ethical Concerns during System Design, in IEEE Std 7000–2021, vol., no., pp.1–82, 15 Sept. 2021, doi: https://doi.org/10.1109/IEEESTD.2021.9536679.

The standard establishes a set of processes by which engineers and technologists can include consideration of ethical values in system design and development.

4. Guszcza, J., Danks, D., Fox, C., Hammond, K., Ho, D., Imas, A., Landay, J., Levi, Ma., Logg, J., Picard, R., Raghavan, M., Stanger, A., Ugolnik, Z., Woolley, A., Hybrid Intelligence: A Paradigm for More Responsible Practice (October 12, 2022). Available at SSRN: https://ssrn.com/abstract=4301478 or https://doi.org/10.2139/ssrn.4301478.

The paper presents the hybrid intelligence paradigm, aimed at supporting a more responsible practice, through simultaneous consideration of machine capabilities and human psychology, behaviors, needs, and values in the development of AI-based systems.

Acknowledgments This work was supported by the Engineering and Physical Sciences Research Council [grant numbers EP/V026747/1 and EP/R013144/1] and Science Foundation Ireland [grant number 13/RC/2094\P2].

References

Barn, B. S. (2016). Do you own a Volkswagen? Values as non-functional requirements. In Human-centered and error-resilient systems development: IFIP WG 13.2/13.5 joint working conference, 6th international conference on human-centered software engineering, HCSE 2016, and 8th international conference on human error, safety, and system development, HESSD 2016, Stockholm, Sweden, August 29-31, 2016, Proceedings 8 (pp. 151–162). Springer.

Bennaceur, A., McCormick, C., Galán, J. G., Perera, C., Smith, A., Zisman, A., & Nuseibeh, B. (2016). Feed me, feed me: An exemplar for engineering adaptive software. In *Proceedings of the 11th international symposium on software engineering for adaptive and self-managing systems* (pp. 89–95).

Bennaceur, A., Hassett, D., Nuseibeh, B., & Zisman, A. (2023). Values@ runtime: An adaptive framework for operationalising values. In *Proceedings of the 45th IEEE/ACM international conference on software engineering—software engineering in society track*.

Cheng, B. H., & Atlee, J. M. (2007). Research directions in requirements engineering. Future of software engineering (FOSE'07), (pp. 285–303).

Cockton, G. (2004). Value-centred HCI. In *Proceedings of the third Nordic conference on human-computer interaction* (pp. 149–160).

Gentile, M. C. (2010). *Giving voice to values: How to speak your mind when you know what's right*. Yale University Press.

Farahani, A., Pasquale, L., Bennaceur, A., Welsh, T., & Nuseibeh, B. (2021). On adaptive fairness in software systems. In 2021 International symposium on software engineering for adaptive and self-managing systems (SEAMS) (pp. 97–103). IEEE.

Ferrario, M. A., Simm, W., Forshaw, S., Gradinar, A., Smith, M. T., & Smith, I., (2016). Values-first SE: Research principles in practice. In *Proceedings of the 38th international conference on software engineering companion* (pp. 553–562).

Hanel, P. H., Vione, K. C., Hahn, U., & Maio, G. R. (2017). Value instantiations: The missing link between values and behavior?. Values and behavior: Taking a cross cultural perspective (pp. 175–190).

Hassett, D., Bennaceur, A., & Nuseibeh, B. (2023). Feel it, code it: Emotional goal modelling for gender-inclusive design. In *Requirements engineering: Foundation for Software Quality: 29th international working conference, REFSQ 2023, Barcelona, Spain, April 17–20, 2023, proceedings* (pp. 324–336). Springer Nature.

Hussain, W., Mougouei, D., & Whittle, J. (2018). Integrating social values into software design patterns. In *Proceedings of the international workshop on software fairness* (pp. 8–14).

Lessig, L., (2000). Code is Law. Harvard Magazine. https://www.harvardmagazine.com/2000/01/code-is-law-html.

Mougouei, D., Perera, H., Hussain, W., Shams, R., & Whittle, J. (2018). Operationalizing human values in software: A research roadmap. In *Proceedings of the 2018 26th ACM joint meeting on European software engineering conference and symposium on the foundations of software engineering* (pp. 780–784).

Nurwidyantoro, A., Shahin, M., Chaudron, M. R., Hussain, W., Shams, R., Perera, H., Oliver, G., & Whittle, J. (2022). Human values in software development artefacts: A case study on issue discussions in three android applications. *Information and Software Technology, 141*, 106731.

Nuseibeh, B., & Easterbrook, S. (2000). Requirements engineering: A roadmap. In *Proceedings of the conference on the future of software engineering* (pp. 35–46).

Perera, H., Hoda, R., Shams, R. A., Nurwidyantoro, A., Shahin, M., Hussain, W., & Whittle, J. (2021). The impact of considering human values during requirements engineering activities. *arXiv Preprint*. 2111.15293.

Robertson, S., & Robertson, J. (2012). *Mastering the requirements process: Getting requirements right*. Addison-Wesley.

Shahin, M., Hussain, W., Nurwidyantoro, A., Perera, H., Shams, R., Grundy, J., & Whittle, J. (2022). Operationalizing human values in software engineering: A survey. *IEEE Access, 10*, 75269–75295.

Schwaber, K., & Beedle, M. (2002). *Agile software development with Scrum*. Prentice Hall.

Thew, S., & Sutcliffe, A. (2018). Value-based requirements engineering: Method and experience. *Requirements Engineering, 23*, 443–464.

Winter, E., Forshaw, S., Hunt, L., & Ferrario, M. A. (2019). Advancing the study of human values in software engineering. In *2019 IEEE/ACM 12th international workshop on cooperative and human aspects of software engineering (CHASE)* (pp. 19–26). IEEE.

Whittle, J. (2019). Is your software valueless? *IEEE Software, 36*(3), 112–115.

Zave, P. (1997). Classification of research efforts in requirements engineering. *ACM Computing Surveys (CSUR), 29*(4), 315–321.

Zave, P., & Jackson, M. (1997). Four dark corners of requirements engineering. *ACM transactions on Software Engineering and Methodology (TOSEM), 6*(1), 1–30.

Digital Humanism Initiative
Technology for Humanity

SPEAKER

Second African Digital Humanism Summer School, Accra, Ghana, 31 August – 05 September 2026

Hans Akkermans

Hans Akkermans is professor of Business Informatics (emeritus) at Amsterdam. He is the Founding Director of the interdisciplinary Network Institute at Amsterdam that studies the interaction between digital technology and society, in which researchers from the faculties of social sciences, humanities, law, economics, and informatics participate. He has worked for many years in knowledge engineering & management, information systems and innovative e-business modelling, with for example applications and innovations in smart electricity distribution networks and the sustainable energy transition that have been internationally field-deployed and are now in industrial and commercial use. His current research interests focus on the interdisciplinary research, education and community service program W4RA (Web alliance for Regreening in Africa, w4ra.org) and education in and on the Digital Transformation (EURIDICE EU project euridice.eu). He holds a cum laude PhD in theoretical physics in the field of nuclear reactions from the University of Groningen.

Lecture 7b: *Software Engineering After AI* (Panel)

Reader article:

Hans Akkermans and Anna Bon: Digital Humanist Education (Extended Abstract). Amsterdam, The Netherlands, August 2026. Extended Abstract submitted to the Digital Humanism – 2nd International Interdisciplinary Research Conference, to be held in Munich, Germany, March 10-12, 2027.

Digital Humanist Education (Extended Abstract)

Hans Akkermans[1] [0009-0003-5982-176X] and Anna Bon[2] [0000-0001-6324-8271]

[1] AKMC Knowledge Management BV, Netherlands, and UDS, Tamale, Ghana
[2] Pangea BV, Amsterdam, The Netherlands
hans.akkermans@akmc.nl
anna.bon@pangea.nl

Abstract. Since the early 1980's a corporate business-like model of education has gradually settled itself in institutions of higher education. It centres around the idea that education is to be delivered as a form of standardized measurable input-output industrial production, with the student in the role of customer/consumer.

The recent advent and uncontrolled spread of (Gen) AI as a kind of educational technology has contradictory affordances and impacts. If wisely used AI may serve as a useful support tool. But it also enables students to outsource their supposed learning to machines to the point that students themselves are learning nothing. Many teachers, researchers and universities are now at a loss how to deal with AI. One may speak of an AI crisis of education.

On the one hand, AI reinforces, further entrenches and even "automates" the business-like input-output approach to higher education. At the same time, we argue, AI is driving this model to its extreme, to the point that it collapses and becomes bankrupt.

More deeply, in current AI there are tendencies to ideologically mainstream a highly distorted picture of what it is to be human, denying the central importance of human relationality. Particularly the notion of AGI is a form of enshittification of the human. New models for education and learning are needed. We discuss, also based on own experiences, projects and experiments, what the characteristics may be of a new model founded on *Bildung* that is fitting to a Digital Humanist approach to education.

Keywords: AI, higher education, managerial efficiency, learning outcomes vs experiences, enshittification of the human.

Note: this is an Extended Abstract
for the DigHum-RES 2 Conference, Munich, Germany, 10-12 March 2027
(version dd 16 Aug 2026, abstract intended for a long paper)

1 Efficiency as Centre of a Mechanized Managerial Model of Education

Since the early 1980's a corporate business-like model of education has gradually settled itself in institutions of higher education: a model centred around the idea that education is to be delivered as a form of industrial production, with an emphasis on standardization and efficiency, with the student in the role of custom-er/consumer (e.g., [1-11]). This model has further entrenched itself via frames of edu-concepts such as objectifiable learning outcomes, measurable rubrics for student assessment, a linear input-output assembly-line conceptualization of the educational process ("delivery") with the student as the object of intervention. Associated EdTech commercial systems for student monitoring (e.g., Canvas, Blackboard), and ever more elaborate administrative frameworks for accreditation have strengthened this trend.

Thus, over the past decades, higher education has increasingly revolved around questions of efficiency: how can public and private education educate large numbers of students at the lowest possible cost while maintaining measurable and comparable standards? Transparency, accountability, and comparability have become important institutional objectives. Universities are evaluated through metrics, rankings, completion rates, student evaluations, and other key performance indicators. European higher education has undergone a process of marketisation and managerialization (i.e., the idea that higher education is to be managed along the lines of private-sector governance models and practices [2-5]). Here, elements of the Bologna Process [6,7], EU benchmarking, OECD indicator systems [10, 11], national performance-based funding regimes and global university rankings have together strengthened an understanding of higher education in terms of measurable outputs, comparability, accountability, efficiency, productivity and competitiveness.

In public higher education, efficiency translates into larger student numbers, higher student-to-teacher ratios, shorter study trajectories, and pressure to graduate students within the nominal duration of their programs. In the Netherlands, for example, both HE institutions and students are encouraged to complete degree programs within prescribed timeframes. Universities have financial considerations to increase throughput, while prolonged study is strongly discouraged. Consequently, both educators and learners are pushed towards making the educational process as efficient as possible.

Digital platforms called Learning Management Systems (LMSs) facilitate the organization and standardization of the complete educational process. Rubrics are used so that assignments can be quantitatively measured and assessed according to predefined criteria, in a quest for consistency and transparency, with the hope to avoid any possible ambiguity and potential dispute about grading. To achieve complete standardization and transparency, combined with efficiency and high throughput, assignments are described as precise as possible. Students must know exactly what is expected of them: the required format, length, numbers of pages and words, structure, content, learning objectives, and assessment criteria, the time spent on it. Ideally, from the perspective

of standardized assessment, assignments should leave no ambiguity about what constitutes a successful answer.

But this is not the real world that students after their studies will encounter, a world filled with real open-ended problems they must learn to address in collaboration with their fellow human beings. The managerialist, mechanistic assembly-line model of higher education leaves students ill-prepared for the real-world problems of the Digital Age and the transformations and societal shocks they bring. New avenues for higher education must be sought and implemented.

2 AI as Culmination – and Breakpoint – of the Mechanization of Education

However, AI is not such a new avenue – despite heavy technology push and marketing, sometimes bordering on mere propaganda (e.g., [12,13]; for an antidote see [14]). It is striking how simplistic, linear and deterministic BigTech background ideas and ideology on technological progress and innovation are, as if nothing is to be learnt from fields such as STS (Science, Technology, Society) social studies, or philosophy of science and technology [15-18].

Instead, (Gen)AI is pushing the managerialist model of education to its logical extreme. Work such as essay writing can be easily outsourced to the computer [19]. The format is known, as is the number of words in an essay, or the structure in which a text must be given. If higher education defines successful learning primarily in terms of producing "knowledge" outputs that satisfy predetermined criteria, GenAI becomes remarkably effective at producing precisely such outputs. But what does the student learn?

Moreover, if efficiency is itself an educational imperative, using AI becomes a rational response. It saves students considerable time and effort. Moreover, the resulting work is often polished, well structured, and convincing. The quality of the output is commonly higher than when students did it all by themselves.

Even when students use a GenAI chatbot creatively, only as a sparring partner with which they interact through thoughtful prompts and iterative dialogue, the nature of their learning process has fundamentally changed. Something important risks disappearing from this process: dealing with difficulty, complexity and failure. There is no moment of failure anymore; there is only success. The end result looks nice. The student or student group can proudly hand in the work that has been done with the help of this extraordinarily capable digital assistant.

Missing in this type of education in the age of Gen AI is the uncertainty, the open-ended messiness of the real world. Missing is that what we experience in the real work

when we encounter a difficulty, ambiguity, conflicting perspectives, incomplete information, failure, and distress. Learning how to respond to such situations is fundamentally different from learning how to produce correct and well-formatted output. What students need to learn is not reproducing prefab knowledge outputs, but acquiring real-world experiences and an associated knowledge-and-action repertoire supporting the capacity for agency in the Digital Age.

3 AGI: The Enshittification of the Human

The practical problems in higher Education triggered by Gen AI expose an underlying problem that is considerably deeper, and much older than a few years. The debate about AI ultimately raises fundamental questions about our conception of the human being. Some visions of Artificial General Intelligence (AGI) suggest that an increasing range of human intellectual activities will eventually be performed by machines. In more radical transhumanist accounts, the human brain itself is understood as a computational system whose functions could ultimately be reproduced technologically. Reasoning, learning, creativity, communication, and decision-making would then no longer be unique human capacities. In this view, as Turing (1950) suggests [20], the human mind can itself be peeled off like an onion: layer after layer, mechanized until, eventually, nothing seems to remain.

This view has been criticized and attacked from many different sides already a long time ago [21-23]. Searle [21] objects, in his Chinese Room thought experiment, that computational production of linguistic behaviour does not establish understanding. Syntax and successful output are not the same as semantics. (And even less so, we would add, if one brings in the perspective of pragmatics, which constitutes an inherently social realm of human communication.) Winograd and Flores [22] go further by arguing that cognition should not be treated solely as information processing. Understanding comes from situated engagement, communication, interpretation, and participation in a shared world.

Digital Humanism also challenges this reductionist conception of human beings (e.g., [24]). Human beings are not merely information-processing systems. Humans are embodied, sentient beings, emotional, social, relational; capable of autonomous and responsible action. Human learning cannot simply be equated with “learning” in the machine-learning sense. The concept of Artificial General Intelligence (AGI) makes this even more clear.

Descartes is often blamed for an undue separation between mind and body. AGI suffers from another undue separation, that of mind and human relationship (community/society). AGI narrows down human intelligence to a set of intellectual tasks. It has to do so in order to lend some credibility to its claims of surpassing human intelligence by machines. What disappears out of sight, then, is that humans are intrinsically embedded in human relational networks of families, communities, societies. Reason is not

just a bunch of intellectual tasks that may be carried out as logical computation; it is rather an inter-human activity that helps understand and regulate human social-relational behaviour (for better or worse). Feenberg, the critical philosopher of technology, calls this "the social life of reason" in the development of technology [16]. From a very different vantage point, this relational, social nature of reason is also well pointed out in African philosophy and ethics that takes community as a starting point rather than the individual [25-27].

Upon ignoring our human relationality, the idea of AGI is based on a solipsistic view on the human individual. The onion peeling-off metaphor is heavily misleading. Upon peeling off the onion there may come off mind-internal processes that lend themselves to logical computation. But what not comes off upon peeling is the whole embedding world of human relationships.

There is an AGI double reductionism at work here. AGI is quick and easy in attributing human qualities to machines, often in a very handwaving fashion (it lacks the careful scientific and experimentally testable approach that Turing strived for). It artificially humanizes machines. But at the same time, by stripping off human relationality and reducing human intelligence to a bunch of intellectual tasks, it dehumanizes humans. Both reductionisms work together in the attempt to promote AGI as a possible and even laudable goal.

Cory Doctorow, the noted sci-fi writer, has recently coined the term enshittification [28]. He refers to the tech business strategy to at first introducing good-quality services "for free" in order to conquer the market. As soon as this has been successful, the same service is degraded in quality and put up for high prices in a semi-monopolistic market. AGI, and AI in general, is at the surface introduced as a so-called neutral, even objective, technology, but underneath it attempts to make mainstream a whole new ideology that degrades what it is to be human. AGI rests on the enshittification of the human.

4 Humanist Learning: A Learner is an Active Social Subject, not an Object of Intervention and Instruction

Generative AI in the higher education system reveals important weaknesses of an increasingly standardized educational process. Such a system may be efficient and formally productive, focused on measurable outputs, but it lacks what education is ultimately intended to achieve. What is at stake is a Humboldtian concept of *Bildung:* the formation of the person through knowledge, experience, reflection, (human) encounter, participation and citizenship in society [29-32].

[Note: in the full paper, here we will elaborate on the Enlightenment "Humboldtian" views on education that will probably be not well known to many modern-day scientific researchers in many disciplines.]

The ongoing wave of digitalization has made this problem even more visible and pressing. Traditionally, the classroom or lecture room was a primary space for learning. Education was centered on acquiring content knowledge. From the 1990s onwards, with the rise of the internet and search engines, access to factual information became increasingly easy. Information was at our fingertips. The importance attached to memorizing content diminished, but the space for the higher-order competences: applying knowledge, understanding, analyzing information, and solving problems was still there, through reading, assignments, practice and academic writing. With GenAI, many more layers of the educational "onion" can be peeled away. Applying knowledge, analyzing information, making judgements, communicating, academic writing, problem-solving (cf. the "Dublin descriptors"), and even aspects of creativity can nowadays be delegated to the machine. Peeling off these layers, what will be left of education? What will make us experts, responsible citizens, reflective practitioners [33]?

Perhaps the foundation of education is precisely what is not expressed as a standardized learning outcome: the lived experience; encountering failure; the joy of achievement after sustained effort; overcoming difficult problems; belonging to a community of learners; exposure to unfamiliar situations and perspectives; dialogue and disagreement; collaboration; and encounters with otherness. These dimensions are currently not included in the general competence-based frameworks for teaching and learning in higher education. Yet, they must become important, precisely because AI systems can perform so many of the activities traditionally regarded as evidence of competence.

We thus argue for an alternative approach, with an emphasis on Bildung. This requires greater autonomy for educators and supervisors, because meaningful educational encounters and experiences must be organized but are difficult to standardize. Universities must then reconsider their reliance on metrics, KPIs, rankings, standardized assessments, optimized throughput, and narrowly defined indicators of student learning. Bildung requires more time and less haste, more reflection, more social interaction and deliberation, and less emphasis on efficiency. Above all, it requires a redefinition of the "skills" that a university graduate is expected to acquire.

From this perspective, education should not be reduced to the acquisition of measurable skills. It concerns experience, understanding, judgment, responsibility, and the formation of individuals competent and capable of participating in society. Higher education should prepare reflective practitioners, who play their part as responsible citizens in an increasingly interconnected world. This dimension is particularly important because the major challenges facing contemporary societies do not respect national borders. Climate change, energy transitions, global health, food and water security, biodiversity loss, and changes in the Earth's dynamic systems are inherently global problems. Addressing them requires not narrow nationalism and chauvinism, but forms of cosmopolitan responsibility, an ability to understand different perspectives, and a willingness to engage with others across cultural, disciplinary, and geographical boundaries. As Seneca says in his Moral letters to Lucilius (Epist. 28): "I am not born for any

one corner of the universe; this whole world is my country." In other words, a "global citizenship" is called for.

[Note: in the full-length paper, to illustrate these points and as an operationalization of our ideas, we will give here a description of the main lines and underlying principles of the EURIDICE EU project curricula and courses in the realm of Digital Society, Social Innovation, and Global Citizenship.]

5 Combatting the AI Crisis in Higher Education

We have argued that Digital Humanism needs an educational dimension: an account of what it means to teach and learn in a humanist, Humboldtian tradition in the age of AI [1]. The emergence of GenAI provides an opportunity to reconsider the purpose of higher education itself. The current challenges posed by GenAI should be used as a catalyst for rethinking higher education. We propose several pathways towards more reflective, experiential, dialogical, and human-centred modes of education — a model that places Bildung at the centre of teaching and learning in the age of AI.

[Note: in the full length paper, we will here summarize (from Secs 2-4) what in our view are salient characteristics of Digital Humanist education, including:

** focus on learners and students as social active subjects, developing their capabilities for agency*

** interdisciplinarity: but in a non-silo form that truly integrates diverse disciplinary perspectives and insights*

** open-endedness: deal with open-ended real-world problems that do not have standardized predetermined answers*

** collaborative: a collaborative approach to problem solving that invites for critical analysis, deliberation, and creative problem solving in a technosocial (but not purely technical) context*

** design of learning experiences, broader than acquiring technical skills and prefab bodies of knowledge.*

** a reflective approach throughout the curriculum, aiming to build not just professionals, but reflective practitioners.*

Acknowledgments. The authors gratefully acknowledge, and this essay has greatly benefitted from, many discussions with colleagues over the years, in particular individuals and partners in the EURIDICE project. The present work has been supported in part by the Digital Europe Programme of the European Commission through the 10 MEUR EURIDICE project (https://euridice.eu), EU Project No. 101123121.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Schmoelz, A., Winter, D., Bon, A., Sabou, M., Hahne, P.-Z.: Teaching for Digital Humanism: Towards a common core? *Journal of Philosophy of Education*, under review (2027). Author version available at
2. Gerhardt, L.-M., Goldenstein, J., Owertel, S., Poschmann, Ph., Walgenbach, P.: The Managerialization of Higher Education in Germany and its Consequences. In Sahlin, K., Eriksson-Zetterquist, U.(eds.): University collegiality and the erosion of faculty authority. Emerald Publishing Limited, Emerald, Leeds, UK (2024), p. 59 - 85. DOI: https://doi.org/10.1108/S0733-558X202386
3. Fisher-Ari, T. R., Kavanagh, K. M., Martin, A.: Urban teachers struggling within and against neoliberal, accountability-era policies. Penn GSE Perspectives on Urban Education, 13(2), 1–19 (2017).
4. Olssen, M.: Neoliberal competition in higher education today: Research, accountability and impact. British Journal of Sociology of Education, 37(1), 129-148 (2016).
5. Hamilton, L. S., Stecher, B. M., Yuan, K.: Standards-based accountability in the United States: Lessons learned and future directions. Education inquiry, 3(2), 149-170 (2012).
6. Štech, S.: The Bologna Process as a New Public Management Tool in Higher Education. Journal of Pedagogy, 2(2), 263–282. DOI: 10.2478/v10159-011-0013-1 (2011).
7. Grunert, C., Ludwig, K.: Academic reform in fractured disciplines– On the Interaction of Bologna, New-Public-Management and the dynamics of disciplinary development. Minerva, 60(1), 57-80 (2022).
8. Schlögl, P.: Digitalisierung – oder: Wenn Technik nicht Probleme löst, sondern erneut eine Bestimmung beruflicher Handlungsfähigkeit herausfordert [Digitisation - or: When technology does not solve problems, but once again challenges a professional capacity to act]. In Dobischat, R., Käpplinger, B., Molzberger, G., Münk, D. (eds): Bildung 2.1 für Arbeit 4.0? (pp. 303–318). Springer VS, Wiesbaden, Germany (2019).
9. Van Berckel Smit, F. J. N.: New Public Management aan de universiteit?: Veelvormige organisatieveranderingen aan Nederlandse en Vlaamse universiteiten, 1970-2010 (PhD Thesis, in Dutch, VU Amsterdam, The Netherlands, 2025).
10. OECD: Benchmarking Higher Education System Performance. OECD Publishing, Paris, France, 2019. https://doi.org/10.1787/be5514d7-en
11. OECD: Education at a Glance 2025: OECD Indicators, OECD Publishing, Paris, France, 2025. https://doi.org/10.1787/1c0d9c79-en
12. Tegmark, M.: Life 3.0 – Being human in the age of Artificial Intelligence. Alfred A. Knopf, New York, NY, USA (2017)
13. Karp, A.C., Zamiska, N.W.: The Technological Republic – Hard Power, Soft Belief and the Future of the West. The Bodley Head, Penguin Random House, London, UK (2025)
14. Narayanan, A., Kapoor, S.: AI snake oil. What Artificial Intelligence can do, what it can't, and how to tell the difference. Princeton University Press, Princeton, NJ, USA (2024)
15. Feenberg, A.: Questioning technology. Routledge, Abingdon, UK (1999)
16. Feenberg, A.: Technosystem – The social life of reason. Harvard University Press, Cambridge, MA, USA (2017)
17. Rogers, E.M.: Diffusion of innovations. Free Press, New York, NY, USA (1st edn. 1962, 5th edn. 2003)
18. Akkermans, H., Bon, A.: Digital Innovation: Myth, Magic, Realities. IWM*post* No. 135, p.18 (Spring/Summer 2025). Open Access available from IWM, Institut für die Wissenschaften vom Menschen / Institute for Human Sciences, Vienna, Austria, https://www.iwm.at/publication/iwmpost-article/digital-innovation-myth-magic-realities

19. Bartlett, J.: How To Talk To AI (And How Not To). Penguin Random House, London, UK (2026)
20. Turing, A.M.: Computing Machinery and Intelligence. Mind, Vol. 59, No. 236 (Oct., 1950), pp. 433-460 (1950)
21. Searle J.R.: Minds, Brains and Programs. Behavioral and Brain Sciences, 3(3), 417-457 (1980).
22. Winograd, T., Flores, F.: Understanding Computers & Cognition: A New Foundation for Design. Norwood, NJ: Alex Publishing Corporation (1986)
23. O'Gieblyn, M.: God, Human, Animal, Machine. Doubleday, New York, NY, USA (2021)
24. Nida-Rümelin, J., Staudacher, K.: Philosophical Foundations of Digital Humanism. In H. Werthner et al. (eds.): Introduction to Digital Humanism – A Textbook, pp. 17-30. Springer Nature, Cham, Switzerland, 2024. https://doi.org/10.1007/978-3-031-45304-5_2
25. Praeg, L.: A Report on Ubuntu. University of KwaZulu-Natal Press, Pietermaritzburg, South Africa (2014)
26. Molefe, M.: Ubuntu Ethics – Human Dignity, Moral Perfectionism, and Needs. Routledge, Abingdon, UK, 2025
27. Metz, T.: A Relational Moral Theory – African Ethics in and beyond the Continent. Oxford University Press, Oxford, UK (2022)
28. Doctorow, C. Enshittification. Verso, New York, NY, USA (2025)
29. Günther, K.H.: Profiles of educators: Wilhelm von Humboldt (1767–1835). Prospects 18, 127–136 (1988). https://doi.org/10.1007/BF02192965
30. Mueller-Vollmer, K., Messling, M.: Wilhelm von Humboldt. In The Stanford Encyclopedia of Philosophy (Winter 2023 Edition), N. Zalta, E.N., Nodelman, U. (eds.). URL = <https://plato.stanford.edu/archives/win2023/entries/wilhelm-humboldt/>.
31. Bacher, S.: Wilhelm von Humboldt's concept of diversity as an integrated component of his idea(l) of Bildung. Ethics and Education, 19:2, 168-184 (2024), DOI: 10.1080/17449642.2024.2364554
32. Esmond, B., Schmees, J.K., Wedekind, V.: Emancipation and Vocational Education: Skills, Bildung and the Subject (1st ed.). Routledge, Abingdon, UK. https://doi.org/10.4324/9781003608141 (2026)
33. Schon, D.A.: The Reflective Practitioner – How Professionals Think in Action. Routledge, London, UK (1983)

Digital Humanism Initiative
Technology for Humanity

SPEAKER

Second African Digital Humanism Summer School, Accra, Ghana, 31 August – 05 September 2026

Saa Dittoh

Saa Dittoh is professor in agricultural development and food systems economy at the University for Development Studies (UDS), Ghana. His interest is in smallholder agriculture, food and nutrition security, irrigation and water management technologies, sustainable natural resources management, ICT for rural development and participatory learning approaches. He is a Distinguished Fellow of the African Association of Agricultural Economists (AAAE). He has worked for international development organizations including UNICEF, WFP, FAO, IFAD, The World Bank, AfDB, World Vision, USAID, DFID, and EU. He has contributed to Ghana's Medium Term Agriculture Sector Development Plans (METASIP I and II), Ghana's Long Term Development Plan (2018 – 2057). Saa Dittoh has been Head of Department, Dean of Faculty, Dean of Students and Pro-Vice Chancellor at the UDS.

Lecture 8: *Integrating Indigenous Knowledge into AI Development in Africa*

Reader article:

Saa Dittoh: The Plug-In Principle: A Theory for Effective Integration of Knowledge Systems and Development Interventions.
In Saa Dittoh, Anna Bon, and Hans Akkermans (Eds.): Integrating Indigenous and Scientific Knowledge for Sustainable Food Systems in Africa. Sustainable Development Goals Series, Springer Nature, Cham, Switzerland, 2025. https://doi.org/10.1007/978-3-031-85512-2

1 The Plug-In Principle: A Theory for Effective Integration of Knowledge Systems and Development Interventions

Saa Dittoh

Abstract

Africa does not lack resources to propel sustainable food and nutrition security for all. The problem has been systematic attempts at replacing supposedly inferior knowledge systems with those thought to be superior. That had been the main cause of lack of sustainability and resilience of development efforts, especially in developing countries. No knowledge system can claim superiority over another. Effective integration of knowledge systems is what will result in the achievement of the SDGs, especially SDG2. The Plug-In Principle describes a methodology for the effective integration of indigenous and "modern" knowledge systems to ensure sustainable outcomes of development interventions. The two-way plug-in, in the right order, ensures increased productivity, resilience and sustainability of agricultural and other systems.

Keywords

Food and nutrition security · Knowledge systems · Indigenous knowledge · Replacement mentality · Effective integration · Plug-In

1.1 Knowledge Systems and the Importance of Integration

Food and nutrition security has been and will continue to be a basic requirement in all countries of the world. Food and nutrition security is necessary in the short, medium and long term. It is not a passing phenomenon; thus, sustainable food and nutrition security is what must be pursued. Unfortunately, in almost all parts of the world, emphasis has been on short term achievements of food and nutrition security. That and other shortcomings of the world development systems gave rise to the need for the millennium development goals (MDGs) and then the sustainable development goals (SDGs) (UNDP and World Bank 2015; Griggs 2015). Developing countries, in particular, have been facing a clash of knowledge systems with respect to food production, distribution and consumption. Should production be small scale or large scale? Should distribution be localized or globalized? Should rice and maize replace millet and sorghum as staple crops? Should roots and tubers be abandoned because they are less responsive to modern production inputs? There have been these and many similar unanswered questions. All the questions need answers and/or clarifications if sustainable food and nutrition security is to be pursued in all countries of the world. The clash of knowledge systems must be resolved. Sustainability, though

S. Dittoh (✉)
University for Development Studies, Tamale, Ghana

S. Dittoh et al. (eds.), *Integrating Indigenous and Scientific Knowledge for Sustainable Food Systems in Africa*, Sustainable Development Goals Series,
https://doi.org/10.1007/978-3-031-85512-2_1

difficult to define adequately, because it is often viewed differently by different persons and under different situations, is valued by all persons. The common view, that it involves meeting the needs of the present without comprising the needs of future generations, is often agreeable to most people even though the actions of many people may be contrary. This means in theory, people of whatever knowledge system or systems they profess, seek sustainability and thus the effective integration of knowledge systems is important. It is the interaction of different knowledge systems that generate new knowledge. According to Buzinde et al. (2020), co-produced solutions (from different knowledge systems) are essential for sustainability outcomes. Achievement of the SDGs, and especially SDG 2, depends critically on the integration of several food and other knowledge systems. Achievement of sustainable and resilient food systems at all levels require integration of knowledge systems at all levels of the food chain: production, post-harvest handling, processing, marketing, consumption and post-consumption. This chapter looks broadly at the integration of knowledge systems and why developing countries, in particular, should avoid replacement, and pursue effective integration, of knowledge systems to ensure sustainability of systems generally and food systems in particular. The neglect of indigenous knowledge systems tends to engender unsustainable outcomes of development interventions.

There have been arguments against the usefulness of indigenous knowledge (IK). Some of the arguments are that "science (often termed Western science) radically differs from all traditional knowledge systems" (Agrawal 2009), and IK and science have "different epistemologies and should not be blended" (Moller et al. 2009); implying that integration of IK and science (Western science) cannot result in any positive outcome. There have however been equally strong arguments for positive outcomes of the integration of IK and science (Veitayaki 1997; Altieri 2004; Berkes 2009; Maclean and Cullen 2009; Athayde et al. 2017). Ludwig (2016), for example, sees knowledge integration to be "often epistemically productive and socially desirable". Manuel-Navarrete et al. (2021), however, argue that to effectively integrate IK and science, there is need "to end the monopoly of science" while Latulippe and Klenk (2020) point out that there is a need to "de-center Western science and institutions as primary sites of knowledge production and leadership". These arguments are not about the relative superiority of knowledge systems but the importance of their effective integration or what Manuel-Navarrete et al. (2021) refer to as "horizontal co-production and decolonization". At the practical level, Dittoh (2003) has given examples of how the integration of IK and science has led to better management of natural resources, improved and sustainable food production practices, and better food preparation and consumption practices in several parts of northern Ghana. There are several other practical examples in the food and agriculture sector that give evidence of positive outcomes of the integration of IK and science (Williams et al. 2020; Bwambale 2023). According to Bwambale (2023), "the recognition of limitations and overlapping strengths between science and indigenous knowledge widely prompts arguments to integrate science and indigenous knowledge". The current rise of agroecological practices worldwide as resilient and sustainable agricultural practices is good evidence of the importance of the integration of IK and science (Altieri 1995, 2004; Pretty et al. 2011; IPES-Food 2018; Chappell et al. 2018).

Knowledge systems have been increasing and improving over centuries and a trace of the trend of some knowledge systems over time can support arguments against over prioritization of some knowledge systems. History and religion (Judaism, Christianity and Islam) has it that civilization started in Mesopotamia at the confluence of the Rivers Tigris and Euphrates (present day Iraq). That means the first knowledge system might have started there (in the garden of Eden).[1] Since then, the world has witnessed thousands, probably millions, of different knowledge sys-

[1]This does not in any way purport to claim that the Garden of Eden is the origin of knowledge. What we do know, and many people accept, is that some knowledge started there.

tems and every one of them started somewhere. That means every knowledge system has been "indigenous" to the people who started it. However, every knowledge system has changed or advanced and had elements that "died" over time. The question that should be asked is what engenders or causes or engineers the start of a knowledge system? The simple answer is a desire to meet a need and/or solve a problem. According to religion, God, at the beginning of man's creation, had a desire to create man and make him have dominion over all things and to undertake agriculture, so He engineered the "garden of Eden" knowledge system, an irrigation knowledge system. That irrigation knowledge system has continued till today with numerous modifications. It is interesting that the Rivers Tigris and Euphrates and their tributaries are still important sources of water for irrigation in Iraq, Turkey, Syria and other places till date. The irrigation knowledge system continued to be very relevant in early times. The River Nile was and continues to be a major source of water for agriculture in several African countries. The people of old in the Middle East always found sanctuary in Egypt because of the agricultural produce of the River Nile.

Other knowledge systems, such as gathering and hunting, also emerged in areas where crop and animal production knowledge systems were not yet known. The hunting and gathering arose out of the need for the people to survive. Those systems also exist till date, though very marginally, because improved systems that can meet the same objectives have emerged. So, in most societies the hunting and gathering knowledge systems have either been completely replaced or new systems have been integrated into them. The hunting knowledge system has turned into a game in several societies, but gathering still exists in areas where some crops/trees and animals have not been domesticated. Products from shea (*Vitellaria paradoxa*), dawadawa (*Parkia biglobosa*), baobab (*Adansonia digitata*) and other semi-wild trees and animals as well as some indigenous vegetables, for example, are still being gathered in Africa because they have not been effectively domesticated. The defense knowledge system using available materials evolved with the hunting knowledge system since people had to defend themselves against wild animals. The simple defense knowledge system has now been developed into a very sophisticated system capable of destroying the whole world, but it is good to recognize its simple origin.

The present-day medical knowledge system also has its origin in the use of locally available medicinal plants to treat diseases. Herbal medicine continues to be important in several societies despite the great advancement in medical practice. The recent COVID 19 pandemic has regenerated great interest in medicinal plants in Africa and other places. The present sophisticated engineering knowledge systems also grew out of the knowledge of using available materials (of wood and iron) to produce tools to lessen human effort used in the production of food and other necessities.

The above discussion is to point out that every knowledge system has an indigenous origin. Thus "people's science" or "ethno-science", that is indigenous knowledge, is the origin of every knowledge system. That is why indigenous knowledge systems do not die. They are "accepted lived experiences". They have proved to be resilient. Their "souls" are imbedded in communities in the modern systems. The "souls" of irrigation in the garden of Eden, of gathering and hunting, of use of bows and arrows for defense, herbs for curing diseases, medieval wood and iron technology for food production etc. are still existing despite massive "modernization". It has been argued that the current sophisticated ICT system with AI originated from indigenous virtual forms of sending messages between households and communities by high-pitched voices and/or loud sounding talking drums that gave coded information that were understood by those who had the knowledge and/or skills (Dittoh 2021). Also, according to Mapara (2009) knowledge forms (or systems) in Zimbabwe have failed to die despite racial and colonial onslaught over time. Indigenous knowledge systems are only resurfaced and transformed in the face of modern contradictions.

To simplify the understanding of the integration of knowledge systems, they (that is, knowledge systems) may be broadly divided into indigenous and "modern" knowledge systems (IK and MK). All modern knowledge (MK) systems are, however, derived from, or are modifications, of indigenous (IK) systems. Indigenous knowledge systems are bodies of knowledge of indigenous people of particular geographical areas that have survived over a long period of time. They are forms of knowledge and knowledge practices that have originated locally and naturally (Altieri 1995). As explained earlier, every knowledge system has been indigenous to the people who started it. However, knowledge systems have changed or advanced over time.

Dittoh (1981) described how agriculture in England (from late 1700s), the US (from late 1800s), Japan, the then USSR and China all evolved from the people (i.e. IK) and how the successes of the green revolution in the Indian sub-continent was short-lived (mainly because it did not evolve from the people). Lester Brown of the American Overseas Development Council who was deeply involved in the Asian green revolution programme had to agree that the green revolution did not represent a solution to the food problem; it was a means of buying time (Brown 1975). Several researchers led by Robert Chambers have since the July 1987 "Farmer First" workshop at the Institute of Development Studies (IDS), Sussex, United Kingdom (Chambers et al. 1989), advocated strong integration of farmers' knowledge (IK) and MK for sustainable agricultural development. Later the Information Centre for Low Eternal Input Agriculture (ILEIA), established in 1984 in the Netherlands, spearheaded the low external input and sustainable agriculture (LEISA) concept which largely promoted IK and its integration with other knowledge systems. ILEIA posited that "Green Revolution technology was neither sustainable nor feasible for many small-scale farmers around the world". No wonder that in Africa, "despite generations of western influence, the decisions about agriculture, health and nature management are still heavily based on the concepts of African traditions" (Millar and Dittoh 2004). The basis for work done by researchers at the Vrije Universiteit, Amsterdam and four African teams in North, West, East and Southern Africa between 2013 and 2016 was that "there are local community innovations succeeding where formal research recommendations have often failed" (Mudhura et al. 2016). Current research continues to point out that lack of integration of IK and MK makes education and research "too academic and distant from the developmental challenges of African local communities" (Kaya and Seleti 2013; Offei et al. 2010; Teeken et al. 2012; Mokuwa et al. 2013). Teeken et al. (2012) showed that ecological and socioeconomic factors as well as cultural norms and values, shape the use and development of local rice technologies, while Mokuwa et al. (2013) showed that West Africa farmers' rice varieties are robust and tolerant of sub-optimal conditions. Navarro et al. (1994) have also argued that hermetic storage[2] is not a new concept and has been recorded from prehistoric times when grain was stored underground in pits. They argue that hermetic bags are an adaptation of the traditional and other forms of hermetic storage systems such as underground pits, clay pots, jerrycans, silos, or drums.

The argument is that sustainable development processes must be based on effective integration of knowledge systems. If MK systems grew out from their indigenous origins because of the need to solve specific problems of societies, it makes sense that the best way for two knowledge systems from different environments to be useful to societies is for their effective integration. The integration should, however, be a two-way process in the right order, to ensure resilience, sustainability and long-term productivity.

[2]Hermetic storage is being promoted for small farmers in Africa and elsewhere mainly because it is a chemical-free, safe and simple means of storage of all types of dry grains (cereals and legumes).

1.2 The Replacement Mentality of Modernization

Despite the scientific findings of the importance of IK in sustainable agricultural production, "development models" being pursued in Africa and elsewhere currently, by Governments, development partners (DPs) and several NGOs, typically try to ignore IK. Development interventions in Africa continue to insist on REPLACING IK systems with "modern" systems (generally termed technology transfer). Several African countries have over several decades drawn programmes and instituted projects to "modernize" their agricultural and other sectors. World Bank assisted Agricultural Development Projects (ADPs) in Nigeria and Nigeria's River Basin Development Authorities (RBDAs), Ghana's Regional Development Corporations (RDCs), as well as several Africa Development Bank (AfDB) and Alliance for Green Revolution in Africa (AGRA) assisted agricultural projects in all parts of Africa are examples of such.[3] What most of the programmes and projects have attempted for several decades is to "replace" the practices of the people, and for decades there have been consistent and persistent failures. Several reports and research papers have continued to bemoan the non-adoption and low adoption of modern technologies and have put the blame on practitioners, mainly farmers (Feder et al. 1981; Acheampong et al. 2018; Mapemba et al. 2019; Kebebe 2019; Balehegna et al. 2020; Kazembe 2021). This process of blaming people (especially farmers) for resisting unproven and indeed destructive modern technologies has been very unfair. It has been based on a strange mentality of "the know-all" teaching the "know-nothing" and the latter should have no say in what the former believe is good for them. As stated by Horowitz (1990), "development interventions (in Africa) designed to enhance the habitat and improve the income and productivity of small-scale rural households so often have exactly contrary outcomes". It is the reason the people do not willingly adopt many of the so-called modern technologies.

Most of the attempts at modernizing Africa's agriculture are watered-down versions of the barely successful Asian green revolution;[4] and so far, there are hardly any evidence of significant achievements (Dawson 2016; Dittoh and Akuriba 2018; Bassermann and Urhahn 2020; Wise 2020). In Africa's agricultural sector, attempts have been made to REPLACE (not improve) the hoe and cutlass, seeds, breeds, indigenous methods of soil amendments, farmer-led irrigation systems, indigenous soil and water conservation systems, land tenure systems, cropping systems, livestock production systems, indigenous agro-forestry systems etc. and hardly any of the replacement attempts can claim any sustainable success. According to FAO (2016), African farm systems are the least mechanized across all continents. Only 1% of all households across six countries (Ethiopia, Malawi, Niger, Nigeria, Tanzania, and Uganda) own or hire tractors (Sheahan and Barrett 2017). It is so despite the fact that there have been substantial state-led efforts in Africa to promote mechanization during the 1960s and 1970s, often involving public machinery imports (Pingali 2007). Most of those efforts, however, failed due to governance challenges according to Pingali (2007). They actually failed because of the "replacement mentality", and the situation is unlikely to change unless there is a change in that mindset. In recent years, there has been talk of use of "Uber-like" mobile mechanization systems (Daum et al. 2021) but that is still begging the critical problem facing agricultural mechanization in Africa. Daum and Birner (2020) have correctly pointed out that there is need for research areas that will help to find mechanization solutions in Africa that are economically, socially, and environmentally sustainable. It is interesting that the Food and Agriculture Organization (FAO) and the African Union Commission (AUC) said in 2018 that "doubling agricultural productivity and eliminating hunger

[3] See Chap. 2 for discussion on some of these agricultural development projects.

[4] B. H. Farmer has edited a book, "Green Revolution?" (Farmer 1980) which has discussed extensively "the large question-mark" of the Asian Green Revolution.

and malnutrition in Africa by 2025 will be no more than a mirage unless mechanization is accorded utmost importance" (FAO-AUC 2018) and yet there is hardly any direction to mechanization development in Africa in 2025. There is a need to start from the basics and remove the replacement mentality. Similar situations can be seen in the seeds and breeds, extension, irrigation and several other sectors and sub-sectors in the agricultural, industrial, health and even food consumption sectors across Africa. Development of every type must be situated within the culture of the people for it to be relevant (Apusigah 2011). Any development initiative which does not integrate well with the culture of people will result in mal-development, and culture is derived largely from IK.

The replacement mentality in agriculture and possibly other sectors have arisen out of the erroneous belief that "modern" systems are superior and indigenous (or local) systems are inferior and that supposedly superior knowledge systems should replace supposedly inferior ones. That has not worked in many situations. That mentality goes far down to culture and social values where prevailing Western social values are not only considered to be superior to all others but are superior to past Western social values. The mentality continues to stifle development in many societies, including the Western societies themselves, because the supposedly inferior societies try to imitate the "superior" societies by abandoning useful and necessary societal values. Societies that have sought development based on local values have been successful most of the time. Maclean and Cullen (2009) point out that "working with indigenous Australians as research partners (rather than end-users of research) is one pathway for the (effective) integration of indigenous and scientific knowledge". Benneh (2011) has also pointed out that "Japanese agricultural transformation was based on improvement on their native technology (the Meji technology)". That can be said of several other successful development models across the globe.

The continued calls for African and other farmers to adopt modern technology are premised on claims that in the agricultural sector, "modern" systems are more productive, efficient and profitable. Several researchers have, however, questioned those claims. As far back as 1975, Uma Lele undertook an in-depth review of 13 donor-funded projects in seven African countries, and concluded that, "except for a few well-conceived export crop schemes (such as smallholder tea and coffee projects in Kenya), integrated rural development projects, as practiced by donors at the time in Africa, had limited impacts and even bigger questions about their sustainability" (Lele 1975). There has been growing agreement that conventional Green Revolution agricultural intensification strategies based on high yielding crop varieties and inorganic fertilizers cannot achieve the desired outcomes of increasing agricultural productivity sustainably, providing adequate nutrition and avoiding environmental destruction (Farmer 1980; Pingali 2012; Kim et al. 2019). Also crop yield responses to fertilizers have been very low (Dittoh et al. 2012; Kassie et al. 2013; Sheahan and Barrett 2017). One recommendation has been a turn to sustainable agricultural intensification (SAI) (Tilman et al. 2002; Petersen and Snapp 2015; Kim et al. 2019). The goal of SAI is "producing food from the same area of land while reducing the environmental impacts" (Godfray et al. 2010). However, obtaining the right outcomes in food-related issues is complex. It will require drastic changes in the food production, distribution and consumption systems especially in this era of climate change. So many other recommendations since the realization of the destructive effects of climate change have been climate smart agriculture (CSA), conservation agriculture (CA), regenerative agriculture, agroecology and several others. Most of these however have their origin in IK. Agroecological systems, for example, are deeply rooted in the ecological rationale of traditional small-scale agriculture (Toledo 1990; Altieri 2004). There have also been arguments that agroecological systems have fed much of the world population for centuries and continue to feed people in many parts of the planet, especially in developing countries, and hold many of the potential answers to the production and natural resource conservation

challenges affecting today's rural landscapes (Koohafkan and Altieri 2010).

Hazell (2013) also argues that "small farmers are the most economically efficient producers", while others hold the view that family farming which main features consists of the preservation of traditional knowledge, sustainable management of natural resources, the empowerment of women and an economic model based on community and solidarity, is the key element in fighting hunger consistently and effectively (Caritas Europa 2014). Dittoh and Akuriba (2018) have also stated that the overemphasis on short term productivity (yields) and profitability goals has been the bane of African agriculture while Muya (2007) has argued that many modern systems in Africa lack indigenous content and thus "tend to be inadequately sensitive to the developmental challenges of local communities and country". This is not an argument against the usefulness of MK but the importance of integration of IK and MK. The argument is for modern technology to grow out of IK within the environments of the IK, and then probably take the lead since more science would have been imbedded in the MK. Most MK systems do not grow out of IK systems prevailing within their specific environments. They are often developed in different environments and then "tested" for technical suitability in other environments. That methodology cannot result in acceptance and sustainability of the developed technology. Sustainability has economic, environmental as well as socio-cultural dimensions and all the dimensions must be important considerations of a sustainable technology.

1.3 Effective Integration of Knowledge Systems: The Plug-In Principle

Development interventions have taken place in various developing, including African countries, over the years but the results have largely been disappointing. Apart from the Uma Lele study referred to above (Lele 1975), several development interventions especially in the agriculture sectors in Africa have fallen far short of the expected results. In many situations there is hardly a trace of the projects a few years after they end (ACB 2016a, b; Dawson 2016; Dittoh and Akuriba 2018; Bassermann and Urhahn 2020; Wise 2020). It has been mainly because interventionists (development workers) have been trying to FORCE change. The system of replacement has been to engender forced change and that is not the way to go. Change that is sustainable can only be NEGOTIATED between equal and mutually agreeable partners; it cannot be forced. That means development and dissemination of appropriate technologies must be by a "bettering process" and not a replacement or transfer or change process. Sustainable bettering process is what is illustrated by the Plug-In Principle (Fig. 1.1). It illustrates a principle for effective integration of IK and MK to ensure sustainability. The Plug-In Principle is based on the premise that interventions though important cannot replace what already exists. They can only improve upon or "better" them through co-creation and/or co-innovation processes. Thus, interventionists can only be "bettering agents" and not change agents. The principle also hinges on the belief that sustainable development is basically a continuous marriage of ideas or integration of knowledge systems, and every stakeholder has an important role to play. It is also a two-way process where MK grows out of IK and should take over as the lead, for a reverse "plug-in" by IK (Fig. 1.5).

Figure 1.1 depicts an IK pipe (A) which is wider than an MK pipe (B). The B pipe represents the "desired mindset" of an interventionist (development agent or worker). It illustrates that effective integration of knowledge systems must start with a "correct" mindset. The mindset of an interventionist at the beginning should be that what is already happening (IK) is more important or more relevant than what is being plugged-in (MK). Any other mindset will greatly compromise achievements. Such a mindset will prevent top-down intervention tendencies. With such a mindset, the interventionist will recognize the need to study, understand and appreciate what is in A (IK), that is, what the people know or do not

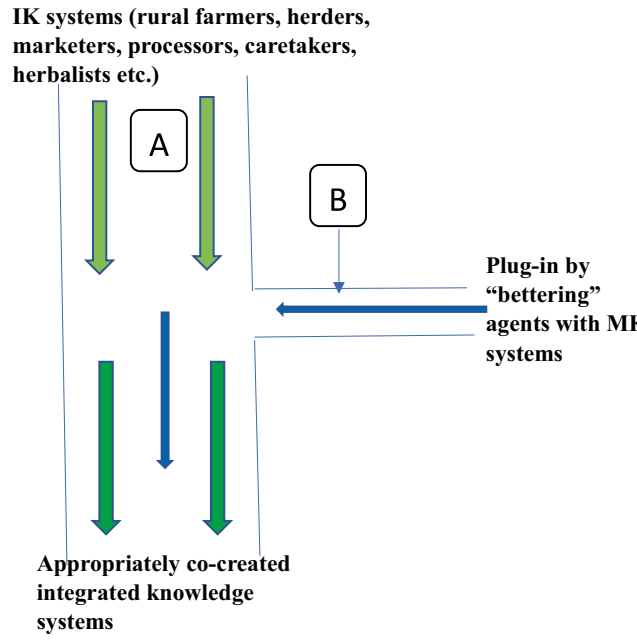


Fig. 1.1 Effective integration process—the plug-in principle

know and their mindset. It is this initial understanding and appreciation of IK and the people (usually called beneficiaries) that will help an appropriate design of a plug-in (MK). Interventionists must realize that the people or community will exist irrespective of what they do or do not do. It means the survival of the people or community is not so much dependent on any intervention. The intervention only "betters" the people's situation. It does not "change" their situation. Thus, development workers, researchers, donors and even governments should consider themselves as "bettering agents". Interventionists (i.e. the "bettering agents") should even see themselves as "intruders" and should bear the onus of success or failure of the interventions. Often, when interventions fail, the "expected beneficiaries" are blamed for non-adoption or non-responsiveness. That narrative must change. Whether an intervention succeeds or fails depends more on what the interventionist does or does not do than on the action or inaction of those already in a process. Sometimes incentives must be given by interventionists to encourage farmers to adopt innovations. However, when a plug-in with incentives must be continuous, then there should be something wrong with it.

Researchers and extension workers also hardly tell farmers the whole truth about "improved" technologies. Farmers, for example, are not usually told the simple truth that produce (grain) from hybrid seeds cannot germinate (they are "dead") or use of fertilizer over time on the same plot can lead to acidity or alkalinity. When farmers find out the truth later, they lose trust in extension messages. Appropriate "plug-in" ensures that the whole truth about a new technology is shared with farmers leading to the building of trust between interventionists and the people. Trust is a critical ingredient in all types of relationships and collaborations.

A measure of sustainability will be the extent to which the "plug-in" can be narrowed without compromising results. The plug-in should continue to get narrower over the years (Fig. 1.2) and ultimately become unnecessary. Successful intervention (plug-in) is obtained if "plug-in" can be narrowed without compromising results. Short-term interventions hardly achieve anything sustainable because they end when it is not yet time to narrow the plug-in. Some end when interventionists are even still trying to understand the IK systems (that is at the A stage).

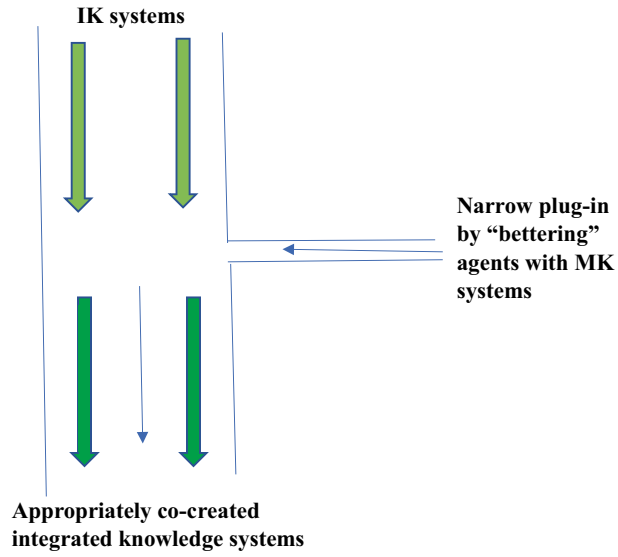


Fig. 1.2 Narrowing of plug-in and positive impact—measure of sustainability of process

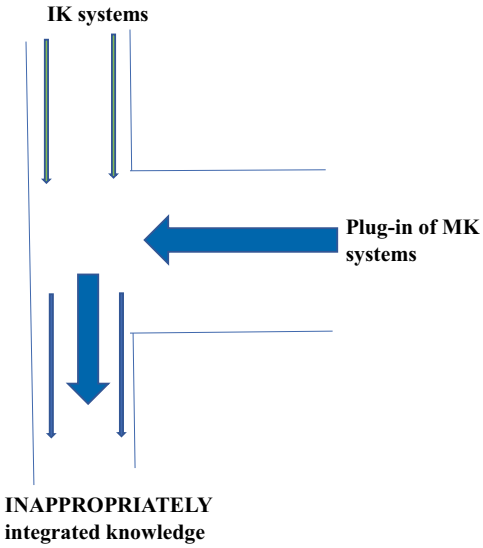


Fig. 1.3 Wrong plug-in—top-down mindset from beginning

Figures 1.3 and 1.4 depict wrong plug-ins, that is, ineffective knowledge integration processes. Figure 1.3 depicts a "top-down" mindset where the interventionist's aim is to "civilize the uncivilized" or the more acceptable term "technology transfer". The concept of technology transfer is a knowledge system replacement, and no society will accept such unless it is forced. As explained earlier, forced interventions are not sustainable.

In Fig. 1.4, the pipe of the MK system is narrower suggesting that the "bettering" agent has a correct mindset, but the actions go contrary to the knowledge and aspirations of the people. May be the interventions go against their traditions, beliefs and/or values. That will also be resisted if not rejected outright.

The process should not, however, end with Plug-In by interventionists (MK). After the co-creation of the "appropriately integrated knowledge system" and the narrowing of the Plug-In (Fig. 1.2), the MK can take lead in the development process since the people would have had a good understanding of the process, and the intervention (MK knowledge system) would have been well integrated with the people's culture. The process would have provided joint understanding of what is possible and works and what does not. MK would now be in a good position to accelerate the development process, but with continuous "plug-in" by IK as shown in Fig. 1.5. The lead by MK at this stage is necessary because the "appropriately integrated system" would now be able to effectively enter the global system. In the area of agriculture, for example, the links to local and international markets, the development of commodity value chains, increases in produc-

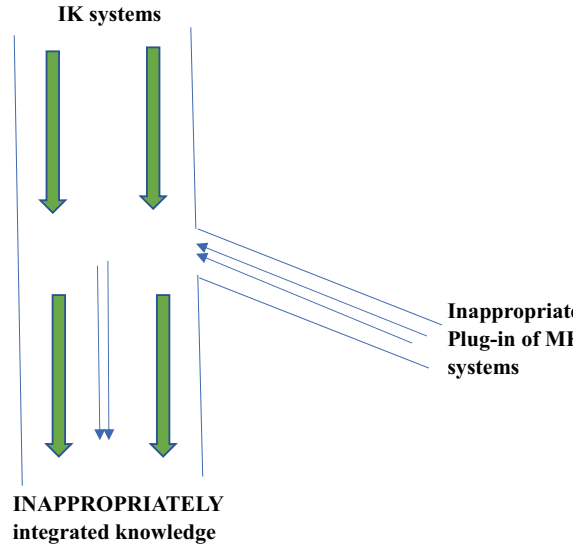


Fig. 1.4 Wrong-plug-in—intervening against the values and practices of the people (Ignoring the importance of IK)

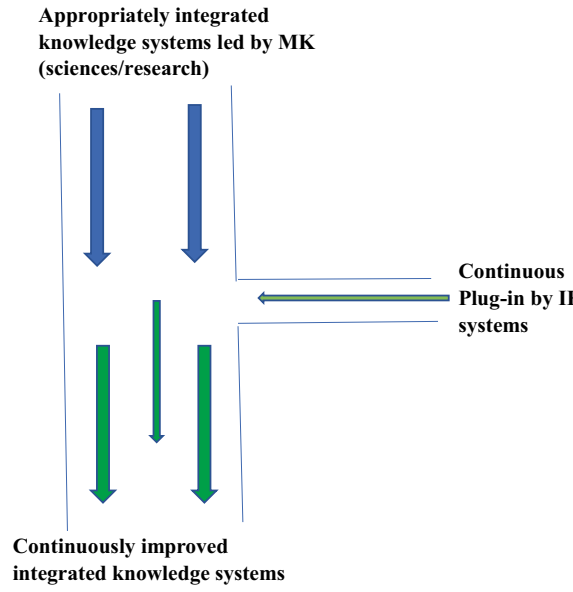


Fig. 1.5 Reverse continuous Plug-In by IK (for resilient, sustainable and productive systems)

tivity and innovations (by the process of co-creation) will at this stage be more effectively driven by MK research (with IK plug-in).

1.4 Conclusion

The persistent African food and nutrition insecurity situation has been difficult to understand. The continent has far more than is required to produce more than enough food for its present and future populations. Africa's uncultivated arable land is about 60% of the world total. Its tropical climate is also probably the best for the production of a wide range of foods. Several parts of Africa can produce two annual crops a year without irrigation. It has agro-ecologies that range from mangrove swamps to semi-arid and arid landscapes and thus almost all types of crops and livestock can thrive. Africa's rivers, the Nile,

Congo, Limpopo, Niger, Volta, Gambia, Senegal and others with their numerous tributaries traverse the continent and groundwater is abundant. There is therefore very little reason for Africa's food and nutrition insecurity. It is instructive that except in conflict areas and severe drought years, most African countries did not experience serious food problems in the 1940s and 1950s. Even presently countries that produce cassava, yam, plantain, maize, sorghum, millet, the pulses, several horticultural crops are self-sufficient in them. The main foodstuffs Africa uses billions of dollars to import are rice, wheat, vegetable oils, meats and sugar. Except for wheat, all other crops can be produced in sufficient quantities even with current consequences of climate change and variability.

The problem therefore seems to lie in the unfortunate replacement of what nature bestowed on Africa (and other continents) with what is considered "modern". Scientific knowledge is indispensable, but it should not try to replace "nature". It should be used to "better" nature and that is the thrust of the SDGs. Johnston and Kilby (1975) argued strongly for "progressive modernization of millions of small-scale farmers" and stated that the starting point of the development process is "the traditional farmer in a traditional economy", while Hazell (2013) pointed out that "the most obvious way to make agricultural growth propoor is to engage with huge numbers of small farms". Those and similar views had been ignored in favour of technology transfer. The technology transfer model of agricultural development has, however, consistently failed because it is based on replacement (change) mentality. What is relevant is effective integration of indigenous and modern knowledge systems—a co-engaged betterment mentality. The Plug-In Principle explains the process of co-engagement and effective "bettering" of nature or indigenous knowledge. Africa and the world will be a better place if the emphasis is bettering nature in the quest to better humanity.

References

ACB (2016a) Green revolution dead-end in Malawi: two case studies—AGRA's Pigeon Pea Project and Malawi's Agro-Dealer Strengthening Programme (MASP). Available at http://acbio.org.za/wp-content/uploads/2016/06/Chinsinga-Report-ACBio-2016-06.pdf

ACB (African Centre for Biodiversity) (2016b) Farm Input Subsidy Programmes (FISPs): a benefit for, or the betrayal of, SADC's small-scale farmers? Available at http://www.db.zs-intern.de/uploads/1468387794-Input-Subsidies-Report-ACBio.pdf

Acheampong P, Amengor NE, Wiredu AN, Frimpong BN (2018) Does awareness influence adoption of agricultural technologies? The case of improved sweet potato varieties in Ghana. Ghana J Agric Econ Agribus (GJAEAB) 2(1):47–60

Agrawal A (2009) Why "indigenous" knowledge? J R Soc N Z 39:157–158

Altieri MA (1995) Agroecology: the science of sustainable agriculture. Westeview Press, Boulder, 433 pp

Altieri MA (2004) Linking ecologists and traditional farmers in the search for sustainable agriculture. Front Ecol Environ 2:35–42

Apusigah AA (2011) Indigenous knowledge, cultural values, and sustainable development in Africa. In: 2nd Annual Ibadan sustainable development summit, Nigeria, pp 1–15

Athayde S, Silva-Lugo J, Schmink M, Kaiabi A, Heckenberger M (2017) Reconnecting art and science for sustainability: learning from indigenous knowledge through participatory action-research in the Amazon. Ecol Soc 22(2):36. https://doi.org/10.5751/ES-09323-220236

Balehegna M, Duncan A, Tolera A, Ayantunde AA, Issa S, Karimou M, Zampaligré N, André K, Gnanda I, Varijakshapanicker P, Kebreab E, Dubeux J, Boote K, Minta M, Feyissa F, Adesogan AT (2020) Improving adoption of technologies and interventions for increasing supply of quality livestock feed in low- and middle-income countries. Glob Food Sec 26:1–11, 100372

Bassermann L, Urhahn J (2020) False promises: the alliance for green revolution in Africa. Biodiversity and Biosafety Association of Kenya (BIBA), Brot für die Welt (Germany), FIAN Germany, and others

Benneh G (2011) Technology should seek tradition – studies on traditional land tenure and smallholder farming systems in Ghana. Ghana Universities Press, Accra, 267 pp

Berkes F (2009) Indigenous ways of knowing and the study of environmental change. J R Soc N Z 39:151–156

Brown L. 1975. World population and food supplies. Agricultural initiatives in the third world. The Agribusiness Council. Lexington Books

Buzinde CN, Manuel-Navarrete D, Swanson T (2020) Coproducing sustainable solutions in indigenous communities through scientific tourism. J Sustain Tour 28(9):1255–1271. https://doi.org/10.1080/09669582.2020.1732993

Bwambale B (2023) Chapter 9: Integrating indigenous knowledge with science to suitably tackle disasters due to climate and environmental change: an overview of the progress and way forward. In: Pal I, Shaw R (eds) Multi-hazard vulnerability and resilience building: cross cutting issues. Elsevier, pp 127–143. https://doi.org/10.1016/B978-0-323-95682-6.00008-5

Caritas Europa (2014) Food security report 2014 – EU's role to end hunger by 2025. Brussels

Chambers R, Pacey A, Thrupp LA (eds) (1989) Farmer first: farmer innovation and agricultural research. Intermediate Technology Publications/The Bootstrap Press, London/New York, xx + 218 pp

Chappell MJ, Bernhart A et al (2018) Agroecology as a pathway towards sustainable food systems. Available at http://www.db.zs-intern.de/uploads/1540849992-synthesis-report-agroecology.pdf. https://doi.org/10.13140/RG.2.2.12122.59842

Daum T, Birner R (2020) Agricultural mechanization in Africa: myths, realities and an emerging research agenda Glob Food Sec 26:1–10, 100393

Daum T, Villalba R, Anidi O, Mayienga SM, Gupta S, Birner R (2021) Uber for tractors? Opportunities and challenges of digital tools for tractor hire in India and Nigeria. World Dev 144:1–15. https://doi.org/10.1016/j.worlddev.2021.105480

Dawson N (2016) Why the 'green revolution' is making farmers poorer in Rwanda. The Conversation, 17 February. Available at http://theconversation.com/why-the-green-revolution-is-making-farmers-poorerin-rwanda-54768

Dittoh S (1981) Green revolution or evolution? The case of independent African countries. Afr Dev VI(3):8–62

Dittoh S (2003) Improving availability of nutritionally adequate and affordable food supplies at community levels in West Africa. In: Brouwer ID, Traore AS, Treche S (eds) Proceedings of the second international workshop on food-based approaches for a healthy nutrition in West Africa. University of Ouagadougou/IRD/Wagenningen University/FAO. University of Ouagadougou Press, pp 51–62

Dittoh S (2021) Knowledge for service – digital technology positives & negatives in African rural societies. Paper presented in DigHum Panel Series, 23 March 2021. Digital Humanism

Dittoh S, Akuriba MA (2018) Africa's looming food and nutrition insecurity crisis – a call for action. Ghana J Agric Econ Agribus 1(1):148–170

Dittoh S, Omotesho O, Belemvire A, Akuriba M, Haider K (2012) Improving the effectiveness, efficiency and sustainability of fertilizer use in sub-Saharan Africa. Policy Research Paper 3, Global Development Network (GDN): The Global Research Capacity Building Program

FAO (2016) Agricultural mechanization: a key input in for sub-Saharan African smallholders. Integr Crop Manag 23

FAO & AUC (2018) Sustainable agricultural mechanization: a framework for Africa. Addis Ababa, 127 pp. Licence: CC BY-NC-SA 3.0 IGO

Farmer BH (ed) (1980) Green revolution? Technology and change in rice-growing areas of Tamil Nadu and Sri Lanka. Macmillan Press, London and Basingstoke

Feder G, Just R, Zilberman D (1981) Adoption of agricultural innovations in developing countries: a survey. World Bank Staff Working Paper No. 44, Washington, DC

Godfray HCJ, Beddington JR, Crute IR, Haddad L, Lawrence D, Muir JF, Pretty J, Robinson S, Thomas SM, Toulmin C (2010) Food security: the challenge of feeding 9 billion people. Science 327(5967):812–818

Griggs D (2015) From MDGs to SDGs – key challenges and opportunities. Future earth research for global sustainability. Monash Sustainability Institute. https://sustainabledevelopment.un.org/content/documents/3490griggs.pdf. Accessed 06/06/2024

Hazell PBR (2013) Options for African agriculture in an era of high food and energy prices. Agric Econ 44(Supplement 1–5):19–27

Horowitz MM (1990) Donors and deserts: the political ecology of destructive development in the Sahel. Afr Environ VII 1–2–3–4(25–26–27–28):185–209

IPES-Food (2018) Breaking away from industrial food systems: seven case studies of agroecological transition. Available at http://www.ipes-food.org/_img/upload/files/CS2_web.pdf

Johnston BF, Kilby P (1975) Agriculture and structural transformation: economic strategies in late-developing countries. Economic development series. Oxford University Press, 475 pp

Kassie M, Jaleta M, Shiferaw B, Mmbando F, Mekuria M (2013) Adoption of interrelated sustainable agricultural practices in smallholder systems: evidence from rural Tanzania. Technol Forecast Soc Chang 80:525–540. https://doi.org/10.1016/j.techfore.2012.08.007

Kaya HO, Seleti YN (2013) African indigenous knowledge systems and relevance of higher education in South Africa. Int Educ J Comp Perspect 12(1):30–44. ISSN 1443-1475 © 2013 www.iejcomparative.org

Kazembe C (2021) The gap between technology awareness and adoption in Sub-Saharan Africa a literature review for the DeSIRA project. IFPRI Project Note

Kebebe E (2019) Bridging technology adoption gaps in livestock sector in Ethiopia: an innovation system perspective. Technol Soc 57:30–37

Kim J, Mason NM, Snapp S, Wu F (2019) Does sustainable intensification of maize production enhance child nutrition? Evidence from rural Tanzania. Agric Econ 50(6):723–934

Koohafkan P, Altieri MA (2010) Globally important agricultural heritage systems: a legacy for the future. UN-FAO, Rome

Latulippe N, Klenk N (2020) Making room and moving over: knowledge co-production, Indigenous knowledge sovereignty and the politics of global environmental change decision-making. Curr Opin Environ Sustain 42:7–14. https://doi.org/10.1016/j.cosust.2019.10.010

Lele U (1975) The design of rural development: lessons from Africa. Johns Hopkins University Press for the World Bank, Baltimore

Ludwig D (2016) Overlapping ontologies and indigenous knowledge. From integration to ontological self-determination. Stud Hist Philos Sci Part A 59:36–45

Maclean K, Cullen L (2009) Research methodologies for the co-production of knowledge for environmental management in Australia. J R Soc N Z 39(4):205–208. 1175-8899 (Online)

Manuel-Navarrete D, Buzinde C, Swanson T (2021) Fostering horizontal knowledge co-production with Indigenous people by leveraging researchers' transdisciplinary intentions. Ecol Soc 26(2):22. https://doi.org/10.5751/ES-12265-260222

Mapara J (2009) Indigenous knowledge systems in Zimbabwe: juxtaposing postcolonial theory. J Pan Afr Stud 3(1)

Mapemba L, Ward P, Bell A, Kenamu E, Nyirenda Z, Msukwa W (2019) Farmer risk perceptions about conservation agriculture: insights from Malawi. In: 2009 AAAE sixth conference, September 23–26, 2019, Abuja, Nigeria, African Association of Agricultural Economists (AAAE)

Millar D, Dittoh S (2004) Interfacing two knowledge systems: local knowledge and science in Africa. Ghana J Dev Stud. FIDS, UDS 1(2):70–84

Mokuwa A, Nuijten E, Okry F, Teeken B, Maat H et al (2013) Robustness and strategies of adaptation among farmer varieties of African rice (Oryza glaberrima) and Asian rice (Oryza sativa) across West Africa. PLoS One 8(3):e34801. https://doi.org/10.1371/journal.pone.0034801

Moller H, Charleton K, Knight B, Lyver O'B (2009) Traditional ecological knowledge and scientific inference of prey availability: harvests of sooty shearwater (Puffinus griseus) chicks by Rakiura Māori. NZ J Zool 36:259–274

Mudhura M, Critchley W, Di Prima S, Dittoh S, Sessay M (eds) (2016) Community innovations in sustainable land management: lessons from the field in Africa. Earthscan studies in natural resource management. University of KwaZulu Natal, South Africa and Routledge (Taylor and Francis Group), London

Muya S (2007) The efficacy of traditional knowledge in African education. Paper presented at the conference on education and Indigenous knowledge systems in Africa. College of Business Education, Dar es Salaam, Tanzania, 2–3 August

Navarro S, Donahaye E, Fishman S (1994) The future of hermetic storage of dry grains in tropical and subtropical climates. In: Highley E, Wright EJ, Banks HJ, Champ BR (eds) Proceedings of the 6th international working conference on stored-product protection, Canberra, Australia, 17–23 April 1994. CAB International, Wallingford, Oxon, pp 130–138

Offei SK, Almekinders C, Crane TA, Hughes SG, Mokuwa A, Nuijten E, Okty F, Struik PC, Teenken B and Richards P (2010) Making better seeds for Africa food security - a new approach to scientist-farmer partnership. *Aspects of Applied Biology* 96

Petersen B, Snapp S (2015) What is sustainable intensification? Views from experts. Land Use Policy 46:1–10

Pingali PL (2007) Agricultural mechanization: adoption patterns and economic impact. Handbook Agric Econ 3:2779–2805. https://doi.org/10.1016/S1574-0072(06)03054-4

Pingali PL (2012) Green revolution: impacts, limits, and the path ahead. Proc Natl Acad Sci USA 109:12302–12308. https://doi.org/10.1073/pnas.0912953109

Pretty J, Toulmin C, Williams SB (2011) Sustainable intensification in African agriculture. Int J Agric Sustain 9(1):5–24

Sheahan M, Barrett CB (2017) Ten striking facts about agricultural input use in Sub-Saharan Africa. Food Policy 67:12–25. https://doi.org/10.1016/j.foodpol.2016.09.010

Teeken B, Nuijten E, Temudo MP, Okry F, Mokuwa A, Struik PC, Richards P (2012) Maintaining or abandoning African rice: lessons for understanding processes of seed innovation. Hum Ecol 40:879–892. https://doi.org/10.1007/s10745-012-9528-x

Tilman D, Cassman KG, Matson PA, Naylor R, Polasky S (2002) Agricultural sustainability and intensive production practices. Nature 418:671–677. https://doi.org/10.1038/nature01014

Toledo VM (1990) The ecological rationality of peasant production. In: Altieri M, Hecht S (eds) Agroecology and small farmer development. CRC Press, Boca Raton, pp 51–58

UNDP and World Bank (2015) Transitioning from MDGs to SDGs. UNDP/World Bank, New York/Washington DC, 176 pp. undp.org. Accessed 06/06/2024

Veitayaki J (1997) Traditional marine resource management practices used in the Pacific Islands: an agenda for change. Ocean Coast Manag 37(1):123–136

Williams PA, Sikutshwa L, Shackleton S (2020) Acknowledging Indigenous and local knowledge to facilitate collaboration in landscape approaches—lessons from a systematic review. Land 9(9):39. https://doi.org/10.3390/land9090331

Wise TA (2020) Failing Africa's farmers: an impact assessment of the alliance for a green revolution in Africa. Available at https://sites.tufts.edu/gdae/files/2020/10/20-01

Digital Humanism Initiative
Technology for Humanity

SPEAKER

Second African Digital Humanism Summer School, Accra, Ghana, 31 August – 05 September 2026

Ricardo Baeza-Yates

Ricardo Baeza-Yates is Search Chief Scientist at You.com as well as part-time WASP Professor at KTH, the Royal Institute of Technology in Stockholm, Sweden. He is also a part-time professor at the Dept. of Engineering of Universitat Pompeu Fabra in Barcelona, Spain, as well as at the Department of Computing Sciences of the University of Chile in Santiago. He is also an affiliated professor at the Data Science and AI division of the Department of Computer Science and Engineering of Chalmers University of Technology and University of Gothenburg as well as at the Data Systems Group of the Cheriton School of Computer Science of the University of Waterloo, Canada. Finally, he is also a visiting professor at the DATA Lab of the Khoury College of Computer Sciences and the Network Science Institute of Northeastern University at its Silicon Valley campus. His research interests includes algorithms and data structures, information retrieval, web search and data mining, and data science and visualization.

Lecture 9: *Responsible AI – Issues and Challenges*

Reader article:

Ricardo Baeza-Yates: An Introduction to Responsible AI. European Review, Vol. 31, No. 4, 406–421, 2023. https://doi.org/10.1017/S1062798723000145

European Review, Vol. 31, No. 4, 406–421

doi:10.1017/S1062798723000145

LECTURE HELD AT THE ACADEMIA EUROPAEA BUILDING BRIDGES CONFERENCE 2022
An Introduction to Responsible AI

RICARDO BAEZA-YATES

Institute for Experiential Artificial Intelligence (EAI), Northeastern University, Boston, Massachusetts 02136, USA Email rbaeza@acm.org

Artificial intelligence (AI) has finally reached most people on our planet thanks to generative AI tools for text and other media. This has started a controversy about the possible benefits and risks, where responsible AI is key. Hence, we introduce the concept of responsible AI, its relation to AI ethics, and why the terms ethical or trustworthy AI should not be used. We then cover the three main relevant aspects of this new field: principles governance and regulation.

Introduction

Artificial intelligence (AI) is increasingly the power behind who we are and who we can be if you are connected to the internet. AI observes how we move, the things we say and the decisions we make. Human beings are frail and susceptible to rash decisions or illness. Our digital selves are what can be improved by AI and, conversely, the ones that could be most at risk from entities that ignore the requirements for responsible AI.

If we ignore the ethical considerations that should be the basis for a functioning society, then there is every chance that AI becomes a cluster bomb as it can reach 5 billion people through the internet: a next-generation tech solution that can be both angel and demon. Ethics, justice and trust are human traits, so it is irrational to assume that the tools we use can either be designed to exhibit the same characteristics or, increasingly, the ethical and fairness considerations that form part of our humanity.

Responsible AI isn't a buzzword. It's a commitment to making the world a better place. From chatbots and voice–user interfaces through autonomous vehicles and the Internet of Things, the ethical questions that AI asks need to be answered, both now and in the future. Responsibility is also a human trait, but our legal system has extended it to institutions, and we use it in that way: an institution should be responsible for bad uses of AI.

In this work, we will discuss the different facets of and our expectations towards responsible AI. This will include but not be limited to what the term means, its history and evolution, the principles behind it and the ways to implement it.

We will also consider the different forms of discrimination and bias in AI systems, the challenges that developers face, the current regulatory landscape and how responsible AI benefits society. Finally, we will look at the current state of AI (including the explosion in generative AI), and the key questions that stakeholders should be asking before they use this technology in their processes.

What is Responsible AI?

With so much media attention and noise currently swirling around the AI sector, it can be easy to forget that the technology hasn't just magically appeared. It has now been 80 years since it first appeared and it became really mainstream a bit more than 10 years ago. But now ChatGPT, launched by OpenAI into the public consciousness in November 2022, has started the equivalent of an AI arms race.

Future historians will likely point to this moment as the catalyst for what came next, but the questions, challenges and implications for widespread AI system usage have been associated with the tech for years. AI has been embedded in our lives for so long that it is often hard to remember when we weren't interacting with it. Amazon's ecommerce dominance was built on AI algorithms, Netflix killed Blockbuster with its state-of-the-art recommendation engines, social media relies on AI to decide what we need to see and when, and big tech wouldn't be investing billions of dollars in pushing AI to the next level if it didn't see a favourable impact on the bottom line.

Before we can discuss responsible AI in detail, we should remind ourselves as to what the term artificial intelligence means. Pop culture fans will always conjure images of sentient computers from a dystopian future, but the reality is less SkyNet and more about automated solutions for mundane tasks.

For the purposes of this article, we are not assessing the state of AI per se. Instead, we are focused on how responsible AI can evolve and, importantly, what its future will (in theory) look like. In addition, we must factor in the concept that machine intelligence is often a mirror of humanity. That means that we need to accept that it can be flawed and prone to biases or learned behaviours.

Simply put, AI is the science and engineering behind the concept of making 'intelligent' machines. The term itself is very broad and covers multiple techniques, including the most popular one, machine learning. That includes neural networks, which when they have many layers, are called deep learning.

https://doi.org/10.1017/S1062798723000145 Published online by Cambridge University Press

In addition, although there is a temptation to think that the technology is actually intelligent, it is unlikely – in its current form, at least – to be plotting the downfall of humanity. Rather, AI tools are developed to perform tasks that seemingly require human intelligence.

The key word here is human. AI relies on data to work effectively, an integral component of every use case that its developers can dream up. In addition, AI is tied to enormous amounts of computational power and financial backing. When you look at AI under these terms, it becomes clear that the technology itself is more akin to a child being taught the difference between right and wrong as opposed to a machine that sees humanity as a threat.

It is very easy to attach convenient terms such as ethical, trustworthy and responsible to a bright and shiny new tech toy, but there are real-world implications of attaching human traits to what is essentially a machine. All three of these words have been associated with AI in the last few years, and there is a danger that the average person will assume they are interchangeable.

Ethical AI implies a degree of moral agency, a human condition that requires value judgements and intended behaviour. However, AI, in its current incarnation, is not a moral agent. Ethics is thereby something that only a human being can demonstrate, so transferring this to a machine is not a good idea.

Trustworthy AI is even more problematic. AI is not human, so we should steer away from attaching humanistic capabilities to not only an algorithm but also a machine that is designed to operate in a vacuum.

We already know that AI does not work all the time, so asking users to trust it is misleading. If, 100 years ago, someone wanted to sell people an airplane ticket calling it trustworthy aviation then the buyer should have been worried, because if something works, why do we need to add 'trustworthy' to it? That is the difference between engineering and alchemy.

Simply put, AI should not be put in either of these buckets. Both 'ethical' and 'trustworthy' are capabilities that can be applied to the human element in the process, but we run into all manner of concerns if they become part of the discourse.

For that reason alone, responsible AI is a better fit. Granted, responsibility is also another human trait, but the concept has the benefit of including ethical considerations and accountability under its umbrella. Institutions are also held responsible for the actions they take, imposing laws and regulations that prevent unethical or illegal behaviour that can harm individuals and the institution itself.

When we think about how wide the responsibility net can be cast, then attaching the right word to a nascent technology is critical. And, from an integration and implementation perspective, the need to define what responsible AI is (and isn't) becomes increasingly important in a society that has been recently upended by the explosion in generative AI as a business optimization and productivity tool.

In a recent interview with ODBMS Industry Watch (Zicari 2022), I said that responsible AI aims to create systems that benefit individuals, societies and the environment. It encompasses all the ethical, legal and technical aspects of developing

and deploying beneficial AI technologies. It includes making sure your AI system does not interfere with human agency, cause harm, discriminate or waste resources.

The caveat is that the words 'responsible AI' have taken on a life of their own. Type them into a search engine and you get 500 million hits in less than a second. Apply a filter for actual news and there are more than 30,000 results generated in the same time period. For context, it takes ChatGPT around three seconds to write a 1000-word blog post on the topic.

There is an increasing number of companies that are trying their hardest to nail down what responsible AI actually is. All of them are likely to be focusing their efforts on the principles behind the design and implementation of the system itself. We will go into these principles in more detail in the third section, but the main ones can be identified as legitimacy, fairness, transparency, accountability, robustness, privacy, security and sustainability.

All of these tick the boxes of what responsible AI can be, but we should think about what that word means. For a human being, responsibility is the moral, legal or mental accountability that we have to each other or an entity or object (Merriam-Webster dictionary). Then, by creating systems that benefit individuals and society as a whole, we must ensure that the system does not interfere with human agency, cause harm, discriminate or waste valuable resources.

Solutions should also be both technologically and ethically robust, a minimum requirement that covers everything from data to algorithms through to design and end user interfaces. In addition, there should always be accountability, not with the system per se but the humans who have control over or input into the AI itself.

The caveat to acknowledging the need for responsible AI is that we must consider what irresponsible AI looks like. Design and implementation are significant elements within the sphere of responsible AI, but even a well-designed system could incorporate illegal or unethical practices.

For instance, developers need to address the consequences of unintended harm, especially when it comes to bias and discrimination. No company is going to admit that their AI is irresponsible, especially if that impacts the brand. But you don't need to be part of the global AI community to understand that machines – much like human beings – can make mistakes. The difference here is that we often judge machines to a higher standard than humans.

Take autonomous vehicles. These cars-of-the-future are packed with AI tools and capabilities. But if one harms or kills somebody, then it is the company that sells the AI that is at fault and not the individual or team that designed that AI to, say, not recognize the pace of an elderly person as they crossed the road or a person of colour in low light. Human beings are involved in thousands of accidents every single day, and it is unlikely that there will ever be a call to ban human drivers from the roads.

AI is also part of our financial system. A bank may decide to run a person's credit through an automated assessment tool to decide whether they should be given a loan. The tool is not looking for reasons to deny that application, but it will draw on every bit of available data to make its decision. If the numbers don't add up, the loan is likely to be denied, irrespective of context and, critically, without a human being

having the final say. Again, we don't apportion blame to the institution but to the automated system that merely did its job.

These are just two examples, but there are literally thousands of others where AI has taken the critical hit, and not the human. If you want to take a deeper dive into why responsible AI matters, the AI Incident Database (https://incidentdatabase.ai/) was set up to index 'the collective history of harms or near harms realized in the real world by the deployment of artificial intelligence systems'. And while that was obviously not the intent of the AI developer, the questions that the widespread adoption of potentially irresponsible AI will bring to the surface are just the tip of the iceberg. AI is currently at a critical inflection point, and the questions that are being asked are not going away.

We will cover the regulatory proposals that are currently being discussed later but it doesn't take much of a stretch to understand that the implementation of responsible AI principles should be at the top of the list. Again, it is people that are building the structures, designing the systems, and training the algorithms, so the onus shifts away from the machine and back to the humans in the loop, although I prefer to say that humans need to be in charge and the computers in the loop.

With that in mind, let's shift our attention to the principles that underlie the concept of responsible AI and the ethics that are intrinsically linked to its required adoption.

Principles and Governance

Before we address the principles and governance required for responsible AI, we should consider the here and now. Algorithmic systems are the backbone of both the digital economy and engrained automated processes, with the public face of AI – ecommerce, streaming services, chatbots and automated decision making – all widely used. This integration is likely to become even more widespread before the end of the decade.

A recent forecast by the International Data Corporation (IDC 2023) said that global spending on AI systems would hit $154 billion in 2023 alone, a compound annual growth rate (CAGR) of 26.9% from the year before. Much of that spending would be on software, hardware and services, with two industries – banking and retail – expected to account for at least 25% of total spend. In addition, IDC analysts noted that only one of the 36 use cases for AI identified within the forecast would have a CAGR of less than 24% over the next five years.

If we unpack this prediction and drill down into what it means for not only AI but also the developers, researchers and, critically, the end stakeholders in this brave new world, then it is clear that we must be very careful about how we both integrate these systems and measure their societal impact. Companies that are slow to adopt AI will be left behind, IDC said, and data-driven decisions would become the norm. That scenario raises the bar on several levels.

Let's not forget that the complexity and opacity of AI systems makes them difficult to understand. When you throw the pace of innovation into the mix,

https://doi.org/10.1017/S1062798723000145 Published online by Cambridge University Press

ensuring that these human-trained models are used in a way that is fair, ethical, and beneficial to society is paramount.

That means that we must adhere to not only a defined set of principles, but also to a form of governance that fits with the concept of responsible AI. In a perfect world, these would align with AI-specific regulations and oversight. We will discuss regulatory activity in more detail in the fourth section, but it will not come as a shock that discussions over how AI can be regulated – a set of rules that will take on added importance in the post-ChatGPT era – can best be described as ongoing.

To fully understand why establishing a set of principles and an effective form of governance is a critical path to take, we should first ask what we want from an algorithmic system. Once we have established our level of expectation, the challenges to responsible AI that these models present can (in theory) become easier to both understand and alleviate.

An algorithm is a self-contained step-by-step set of operations used to perform calculations, data processing, and automated reasoning tasks. The key component is data. For example, most AI algorithms are based on statistical models that are 'learned' or 'trained' from datasets through machine learning (ML). This is called supervised ML. Those that don't fall under this umbrella are driven by analytics, unsupervised ML: the discovery, interpretation, and communication of meaningful patterns in data. The picture is completed with reinforcement learning – learning online – and symbolic AI, among others.

Algorithmic systems have been part of the private and public sector for many years. Companies integrate these tools into their workflow processes to make decisions about potential hires and staff assessment. Banks and financial institutions use them to assess customer spending or credit-based loans, while the digitization of healthcare has allowed algorithms to make judgement calls about patient wellness. Algorithmic decisions can be found in the criminal justice system, often without substantive review by human beings.

From an AI perspective, the increased use of and adoption into society is not breaking news. It should also not come as a surprise that efficiency and decision-based standardization are often cited as the reasons why the machine, and not the human, is the one who passes judgment.

The problem that we face is that while these systems have been designed with the utopian ideal of making society more equitable, inclusive and efficient, automation can also be subject to the same biases and discriminatory practices that are part of the human condition. Much like their human designers, algorithms can fail to respect the rights of individuals and be prone to harmful discrimination or other negative effects.

That makes it even more critical that we establish both guidelines for algorithmic systems to follow and rules that comply with established legal, ethical, and scientific norms. In addition, there is a requirement that the risks associated with their use be proportional to the specific problem that these systems have been designed to address. And while there has been a significant increase in the conversations around what these rules or guidelines should look like, the core requirements of what we should want from an algorithmic system have not changed.

https://doi.org/10.1017/S1062798723000145 Published online by Cambridge University Press

These requirements are not extensive. Common wisdom tells us that they are likely to mirror what we expect from human beings, albeit that a person has agency over decision-making, and a machine is merely adhering to what it has been taught. An algorithmic system, thereby, must have a level of transparency and accountability, an awareness of bias, access and redress, an explanation for its decisions, and date provenance. Additionally, it must display auditability, validation, and be tested in real-world scenarios. However, the algorithms and other underlying mechanisms that artificial intelligence or machine learning systems utilize to make their predictive decisions can be opaque. This has the effect of not only making these decisions less understandable but also difficult to work out whether these outputs are biased or erroneous.

This lack of transparency is an oft-cited challenge in the scientific community, especially when it comes to assessing what impact AI will have on society as a whole. Factors that contribute to these uncertain outcomes can be informational, technical, economic, competitive, and social. AI can produce unexplainable results, and it is an accepted fact that even the most well-engineered algorithmic systems will have bugs or training data that might not align with the intended use.

For instance, the informational data used to train models and create analytics could have been obtained without the data subject's knowledge or explicit consent. The cost of being transparent might not be feasible (from an economic standpoint) or the algorithm itself could be too technical to allow for easy interpretation. In addition, transparency has the potential to hand an advantage to third parties in the form of trade secrets or the ability to game the system itself. All these factors can be influential in how the system is deployed and, crucially, any problems with bias or inaccuracy in the automated decision-making process.

According to the Association for Computing Machinery (ACM 2022), there are nine core principles that developers, system builders and AI policymakers must adhere to. These can be defined as follows:

- Legitimacy and competency
- Minimizing harm
- Security and privacy
- Transparency
- Interpretability and explainability
- Maintainability
- Contestability and auditability
- Accountability and responsibility
- Limiting environmental impacts

All of these core principles were introduced by the ACM Technology Policy Council (TPC) in October 2022, and I was one of the two lead authors. These principles are designed to recognize that algorithmic systems are used by governments, private companies, and other entities to 'make or recommend decisions that have far-reaching effects on individuals, organizations, and society'.

https://doi.org/10.1017/S1062798723000145 Published online by Cambridge University Press

Additionally, the TPC has stated that a 'one size fits all' approach to responsible AI systems is unlikely to be effective, noting that the context of that system will play an integral role in its development and implementation. However, TPC also believes that the establishment of these principles will lead to further discussions among stakeholders, researchers and policymakers. These conversations will both help to chart the course of AI governance by focusing attention on what that will look like and promoting the 'reliability, safety, and responsibility of algorithmic systems'.

The caveat to the establishment of these instrumental principles is that the current state of AI has uncovered significant (and potentially game-changing) challenges. If we take the generic issues that underpin the integration of AI systems into the public sphere out of the equation – principles versus techniques, regulation, our own cognitive biases, and cultural differences – then we find that there are other roadblocks on the path to responsible AI.

These include questions over responsible versus trustworthy AI (see the Introduction), discrimination (facial recognition, justice, the sharing economy, large language models), phrenology (pseudo-scientific biometric-based predictions), unfair digital commerce (end user exposure and popularity bias), and the stupidity of the models themselves due to human incompetence.

The latter is arguably a contributing factor to how we define what it means to be responsible, especially when we are thinking about how systems are used in such a way to be fair and ethical. End users should be able to understand how AI systems work and how they make decisions. At the same time, there is an acknowledgement that these tools are intended to augment and not to replace or harm human beings.

Taking all the above into account, we need to shift our attention to AI governance. The principles set out by the TCP are a good platform to build upon, but the deploying entity must adhere to a clear governance structure when deciding to design and launch an algorithmic system. Applying the principles will lead to necessary trade-offs, and it is expected that these will be dependent on the context of the deployment, the influence of public policy and, increasingly, the market forces that incentivize the development of the systems themselves. In Figure 1, we depict how the ACM principles should be instantiated during the system lifecycle as well as the main tools that need to be used at each stage.

So, what does effective governance entail? And how can we apply a comprehensive and integrated approach to the thorny question of responsible AI itself? The answer to both to these questions can be found in how we approach the subject of ethics in an algorithmic system.

An ethical approach to problem-solving centres on one core concept; what is the right thing to do? The key to solving the issue of responsible AI, according to the puzzle-solving in ethics (PiE) model, comes from not getting lost in the rules, regulations and approvals that come with asking ethical questions. Making the right decision from day one is paramount, as is the need to avoid the problems associated with an undefined strategy or lack of audit processes. These become more integral to the conversation when we factor in the increased power of algorithmic systems and the technical challenges that they present.

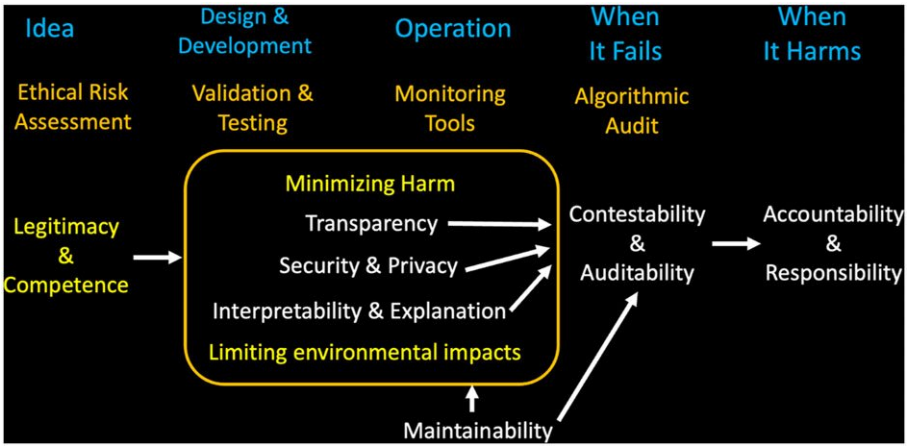


Figure 1. Principles and tools across the lifecycle of an algorithmic system.

At the Institute for Experiential AI (EAI), for example, we use the PiE Model that was developed by the AI Ethics Lab (Canca 2019). This model provides a roadmap to responsible AI that includes governance, the product and the skills required for effective integration. In addition, the institute has created a high-level AI Ethics Advisory Board (Northeastern Global News 2022) and provides deep technical audits of an AI system to determine if there is any bias, appropriate use of data, security and, importantly, the right quality. The last aspect is of vital importance, even more so when you look at how generative AI has been released into the wild.

OpenAI – ChatGPT's developers and owners – has publicly stated that it wants to keep AI 'safe and broadly beneficial', and reportedly spent more than six months working with a more powerful version (GPT-4) to ensure that it ticked as many ethical boxes as possible (Open AI 2023). Real-world use cases are far more difficult to predict than those tested in a laboratory setting, and the company has said that society will take time to adjust to how AI evolves and the potential for harm.

As we have noted before, AI systems are everywhere, but that merely raises the bar in terms of how we promote responsible AI and ensure that everyone is playing by the rules. That means we need to not only regulate AI but also agree on what those regulations are. In the next section, we will discuss why that is not as easy as it sounds.

Regulation

Investment in and integration of AI is top-of-mind for companies of all sizes, but it is the reference points and expectations that have been attached to this next-generation tech that are important to implement in any AI future. Regulation, an essential part of the puzzle, remains a work-in-progress.

https://doi.org/10.1017/S1062798723000145 Published online by Cambridge University Press

As noted in the previous section, acknowledging the principles and intended governance of AI is only the first step. What matters is how we regulate the use of AI in the real world. At time of writing, the regulatory landscape is akin to the Wild West. This situation has not been helped by the emergence of generative AI tools such as ChatGPT and the rush to integrate these large language models into workflows across a wide range of industries.

For the wider AI community, this very public awareness of what AI systems are capable of has raised not only questions about how much people know about the technology but also what level of regulation will be required to keep it under control.

ChatGPT, for example, became the fastest-growing consumer application ever, attracting more than 100 million Monthly Active Users (MAU) in the initial eight weeks following its release in November 2022. Other generative AI systems were quickly released by big tech companies, with an almost inevitable level of regulatory interest from global government agencies following soon after.

Reuters reported (Shepardson and Bartz 2023) that the US National Telecommunications and Information Administration (NTIA) was curious to know if regulatory measures could be applied that 'provide assurance that AI systems are legal, effective, ethical, safe, and otherwise trustworthy'. The Biden Administration, the news source noted, was concerned about whether these tools were 'dangerous', with the NTIA expected to draft a report that would inform the administration as to AI's present and potential future. This cautious approach is not a surprise, especially from a US agency that advises the President on policy direction.

The key point to remember is that the United States is not acting in a vacuum and there will be similar initiatives in progress in other countries.

To date, the number of actual regulatory processes that are ready to be signed is non-existent. Current AI regulations in the US and Europe, for example, fall into two camps: laws that could apply to the use of AI systems and proposed regulatory frameworks or blueprints for the future.

The US does not have any specific AI regulations, but the White House Office of Science and Technology Policy (OSTP) published a 'Blueprint for an AI Bill of Rights' (White House 2022) in October 2022 that establishes five principles for the design, use and deployment of what the agency refers to as 'automated systems'. According to the OSTP, this framework is a response to the concerns of public citizens and will be a significant step towards protecting people from the (unnamed) threats of artificial intelligence.

The principles and the associated practices are exactly what you would expect from an exploratory guideline – safe and effective systems, algorithmic discrimination protections, data privacy, notice and explanation, human alternatives, consideration, and fall-back. What is more interesting is that the federal government is not only pushing for a defined policy for AI – with the very public backing of President Biden – but also drawing on laws or regulations that could apply to AI systems.

For instance, the Federal Trade Commission Act (FTC Act) prohibits unfair or deceptive trade practices. This could potentially include using AI systems in a way

that is deceptive or harmful to consumers. In addition, the Cybersecurity and Infrastructure Security Agency (CISA) has published several guidance documents on how to secure AI systems. Both approaches are still nascent in terms of regulating AI in its current incarnation, but they do provide some evidence that algorithmic systems are likely to become regulated sooner rather than later. The US Chamber of Commerce recently released a report that called for discussions around the use of AI to be 'ramped up', noting that the technology had the capacity to hurt economic growth or become a national security risk – a stance that goes against its usual anti-regulation rhetoric (US CoC 2023).

Europe, on the other hand, has adopted a more strict approach to AI regulations. The EU's General Data Protection Regulation (GDPR) became enforceable in 2018, governing the way in which public companies process personal data that relate to an individual. This legislation adheres to a set of defined principles that include fair and lawful processing, purpose limitation, and the concept of data minimization and retention. GDPR was designed to protect our personal data, with the law stating that individuals must be informed about how their data are used and their data protection rights.

In the five years since it came into effect, GDPR has changed the way that data are viewed, even more so when the data become the subject of a breach or unauthorized use. And while the law doesn't specifically mention AI or algorithmic systems, there is a tacit understanding that they could fall under GDPR's Article 22. Automated decision making has become more of a concern in recent years, and it is expected that the new wave of generative AI will increase conversations over whether the provisions of GDPR can be brought into the regulatory mix.

From a pure AI standpoint, however, the European Union has been proactive. In 2021, the EU published a proposal for an Artificial Intelligence Act (AIA). If adopted, the AIA would create a comprehensive regulatory framework for AI systems and classify them into four risk categories:

- *Unacceptable risk*: AI systems that pose a serious threat to safety, fundamental rights, or public security would be prohibited.
- *High risk*: AI systems that pose a significant risk to safety, fundamental rights, or public security would be subject to strict requirements, such as mandatory conformity assessment and market surveillance.
- *Limited risk*: AI systems that pose a limited risk would be subject to certain requirements, such as transparency and documentation.
- *Minimal risk*: AI systems that pose a minimal risk would not be subject to any specific requirements.

This proposal has been making its way through the requisite regulatory processes since it was introduced and, while it is seen as a positive step forward, the attempts by the EU to address the various elements of AI within a standardized framework have (at time of writing) not produced a concrete piece of legislation. Granted, the EU is focused on risk management, but it can be argued that the AIA is the first law that

deals specifically with the impact that AI will have on public health, safety, and well-being.

High risk systems, for example, would fall under the most stringent of regulations. This would include any AI that is deemed to be a safety component of a regulated product – medical devices, machinery, hazard detection in motor vehicles, to name just three. Biometric identification and categorization of natural persons by law enforcement would also be in this category, with developers of AI systems required to establish the intended purpose of the algorithms.

In addition, any AI regulations would have to consider the end user and the outcomes that the AI produces – conversational AI is, for instance, a part of child development with millions of children interacting with voice–user interfaces such as Amazon's Alexa or Google Assistant. When you factor in a familiarity with recommendation systems via child-friendly apps, then the risks associated with AI take on a new level.

According to the lawmakers behind the legislation, the EU AI Act (European Union 2021) could become a global standard. There is a need to determine the extent to which AI can be a positive as opposed to a negative influence on society, and it is already reportedly providing a framework for other countries to consider their own regulatory stance on AI. The caveat is that there are some identified flaws with the AIA, including the fact that it is (in its current form) relatively inflexible and reflects the point in time in which it was first drafted.

In my opinion, there are three significant problems with the proposed EU laws. First, we should not be looking to regulate the use of the technology but focus on the problems and sectors that the technology impacts. There is an element of fear around the potential for human harm, but we should be approaching AI in the same way that we enact oversight on food or healthcare. In other words, we have to draft regulations that work for all possible technologies and not just the versions that exist in the here and now.

Second, we must understand that risk itself is not static. Rather, it exists as a continuous variable. Companies that choose to integrate AI into their business practices are unlikely to want to allocate it to the four risk categories proposed by the AIA, and there will be the potential for a conflict of interest in terms of self-evaluation. A better way of looking at the potential risks is to compare AI regulation to improvement mechanisms such as the US Food and Drug Agency – an accepted path that takes a phased approach to the approval of a medicine or device through study and human validation.

The third and final problem that the proposed regulations don't address is the issue of systems that do not use AI. Not all algorithmic systems are human-trained, rather they rely on interpreting the data that they are presented with and decide based on the information itself. The use of automated decision-making tools is extremely prevalent in sectors such as recruitment and finance, and there is an argument to be made that the EU AIA Act should apply to all algorithmic systems, not only AI-based systems. Otherwise, there is an easy loophole to skip regulation.

This becomes even more of a concern when we come back to the influence of generative AI. Large language models are the face of AI in 2023, and there is already a backlash against tools such as ChatGPT and its presumed successors. Large language models are being promoted as productivity tools, with companies already using them to automate customer service, generate reports, create content, streamline sales or marketing, and more (Leighton 2023).

This increase in AI usage and awareness has now forced various governmental agencies to fast-track discussions about what regulatory actions will be needed. Notwithstanding the aforementioned EU Laws, the *Financial Times* reported that the EU is set to introduce a 'sweeping set of regulations on the use of AI', with developers required to be more transparent on how these models are trained, and whether they infringe copyright (Espinoza and Johnston 2023).

Italy actually banned ChatGPT, citing a breach of strict privacy regulations. That ban lasted a month, and the tool was only restored after OpenAI agreed to make the changes required by Italian regulators. Germany and France have taken a less gung-ho approach, albeit that both countries are waiting to see how other EU members are dealing with the increase in AI usage; Spain, and Ireland, for instance, have decided to launch investigations into its usage and potential harm, but these are not expected to mirror the Italian approach.

No longer beholden to the EU's decision-making process, the United Kingdom government recently released a white paper (UK Government 2023) that sets out proposals for regulatory activity, although its authors were careful to note that the principles under discussion were more akin to practical guidance as opposed to concrete rules around AI use. The UK is also planning to both review the AI market and assess the potential for 'business domination' by a single tech entity or vendor. According to BBC News (Thomas 2023), the Competition and Markets Authority (CMA) is concerned that two of the companies aligned with the current AI revolution – Microsoft and Google – could have a significant localized impact, with the software behind large language models (LLMs) having the power to 'transform the way businesses compete as well as drive significant economic growth'.

Away from Europe, we must not overlook how China and the US are dealing with this wave of innovation and adoption. China does not have access to ChatGPT, and its draconian internet censorship laws could be a blessing in disguise from a regulatory standpoint. The Chinese Communist Party is considering measures (Digichina 2023) for the 'Management of Generative Artificial Intelligence Services' to fall in line with the country's existing Cybersecurity Law, so it is safe to assume that regulations to govern the use and commercial development of AI are not far away. Companies such as Alibaba, Huawei, Baidu and Tencent are all working on generative AI, and it will be interesting to see how and when they decide to integrate these experiments into their products.

In the US, the Biden Administration is keeping a close eye on developments but has taken no actual action (to date) towards investigating how the tools are being used or introducing regulations. However, the White House did announce (White House 2023) a series of actions that will 'promote responsible AI innovation', which

included a $140 million investment into seven National AI Research Institutes and a public assessment of existing generative AI systems. The latter included a meeting on 4 May 2023 between President Biden and the CEOs of Anthropic, Google, Hugging Face, Microsoft, NVIDIA, OpenAI, and Stability AI, all of whom reportedly agreed to an evaluation of their systems as part of the Blueprint for an AI Bill of Rights and AI Risk Management Framework.

The Brookings Institute reported (Engler 2023) that the 'U.S. federal government's approach to AI risk management can broadly be characterized as risk-based, sector-specific, and highly distributed'. According to the institute, any advantages to this approach can be negated by the fact that it contributes to a perception of 'uneven development of AI policies', with the authors of the report noting that the publication of several guiding documents has not led to the required AI regulations.

At the beginning of this section, we noted that AI regulation is a work-in-progress. And while the discussions around what it should look like are gathering pace, the current fascination with and deployment of algorithmic systems make regulation less of a talking point and more a required action. The question that we need to ask, however, is what timeline are we hoping for?

The EU Parliament just approved the new version of the EU AI Act in early May 2023 (including an article on generative AI) and expects to sign AIA into law in 2024, but there are no guarantees that global regulation will be enacted anytime soon. The challenge, thereby, is not whether we need regulation (which we do), but what that regulation will look like and how it will affect both the need for responsible AI and the principles that underlie its implementation. Until we know the answers to this simple question, then the analogy of the Wild West is likely to be in play for the foreseeable future.

Conclusions

Mark Weiser, the acknowledged father of ubiquitous computing, once said that a good tool is an invisible tool. His analogy was that the blind man tapping the cane feels the street, not the cane itself. In other words, it must be something that does not intrude on your consciousness and, by association, is not there to cause you harm. AI can be that tool, but the human element is what gives it responsible and ethical guidelines.

Science fiction has arguably become our current reality. The global village predicted by Marshall McLuhan back in the 1960s is no longer a vision of the future, it has become part of our connected present, a digital ecosystem that touches every part of our lives. Devices that we use daily have augmented our humanity, an evolution that provides us with access to a second brain whenever we need it. These connected devices don't just define who we are, they present a digital self. In addition, automated systems that have been designed and integrated by human beings to not only alleviate mundane tasks, but also allow companies to be more streamlined and efficient are ubiquitous in every industry sector.

https://doi.org/10.1017/S1062798723000145 Published online by Cambridge University Press

When these two factors are combined, they create a virtual footprint that is intended to both enhance the lives that we lead and the work that we do. The caveat is that the tools that we use and the solutions they provide are, essentially, unthinking systems. And while they may have been developed with good intentions, the unintended ethical consequences of machines programmed by humans could be a real and present danger.

Ultimately, responsible AI should not be an afterthought, but until we have the regulatory guidelines and guard rails in place, AI systems will need to be integrated in such a way that all stakeholders can feel comfortable with how they are being used and, importantly, their impact on society. AI can be a game-changer on so many levels, but all games need to have rules. And the players must follow them. If they don't, then this bright and shiny new tech tool has the capacity to impact our lives in ways that we will find difficult to predict or even plan for.

Acknowledgements

Thanks to David Bolton for his English revision of this article.

References

ACM (2022) *Principles for Responsible Algorithmic Systems*. Lead authors: Baeza-Yates R and Matthews J. https://www.acm.org/articles/bulletins/2022/november/tpc-statement-responsible-algorithmic-systems

Canca C (2019) *The PiE Model: Our Ethics Model for Innovation*. AI Ethics Lab. https://aiethicslab.com/pie-model/

Digichina (2023) *Translation: Measures for the Management of Generative Artificial Intelligence Services*. Stanford University. https://digichina.stanford.edu/work/translation-measures-for-the-management-of-generative-artificial-intelligence-services-draft-for-comment-april-2023/

Engler A (2023) The EU and U.S. diverge on AI regulation: A transatlantic comparison and steps to alignment. Brookings Institute. https://www.brookings.edu/research/the-eu-and-us-diverge-on-ai-regulation-a-transatlantic-comparison-and-steps-to-alignment/

European Union (2021) *The AI Act*. https://artificialintelligenceact.eu/

Espinoza J and Johnston I (2023) European Parliament prepares tough measures over use of AI. *Financial Times*. https://www.ft.com/content/addb5a77-9ad0-4fea-8ffb-8e2ae250a95a

IDC (2023) Worldwide spending on AI-centric systems forecast to reach $154 billion in 2023. https://www.idc.com/getdoc.jsp?containerId=prUS50454123

Leighton N (2023) 6 ways business leaders should integrate ChatGPT. Forbes. https://www.forbes.com/sites/forbescoachescouncil/2023/02/22/6-ways-business-leaders-should-integrate-chatgpt/?sh=535371326c61

Northeastern Global News (2022) Northeastern launches AI Ethics Advisory Board to help chart a responsible future in artificial intelligence. https://news.northeastern.edu/2022/07/28/ai-ethics-board/

Open AI (2023) Our approach to AI safety. https://openai.com/blog/our-approach-to-ai-safety

https://doi.org/10.1017/S1062798723000145 Published online by Cambridge University Press

Shepardson D and Bartz D (2023) US begins study of possible rules to regulate AI like ChatGPT. Reuters. https://www.reuters.com/technology/us-begins-study-possible-rules-regulate-ai-like-chatgpt-2023-04-11/

Thomas D (2023) AI investigated in UK over fears of domination. BBC News. https://www.bbc.com/news/business-65478156

UK Government (2023) AI regulation: a pro-innovation approach. Department for Science, Innovation and Technology and Office for Artificial Intelligence. https://www.gov.uk/government/publications/ai-regulation-a-pro-innovation-approach

US Chamber of Commerce (2023) Report of the Commission on Artificial Intelligence, Competitiveness, Inclusion, and Innovation. https://www.uschamber.com/assets/documents/CTEC_AICommission2023_Exec-Summary.pdf

White House (2022) Blueprint for an AI Bill of Rights. https://www.whitehouse.gov/ostp/ai-bill-of-rights/

White House (2023) Biden–Harris administration announces new actions to promote responsible AI innovation that protects Americans' rights and safety. https://www.whitehouse.gov/briefing-room/statements-releases/2023/05/04/fact-sheet-biden-harris-administration-announces-new-actions-to-promote-responsible-ai-innovation-that-protects-americans-rights-and-safety/

Zicari R (2022) ODBMS Industry Watch: on responsible AI. Interview with Ricardo Baeza-Yates. https://www.odbms.org/blog/2022/02/on-responsible-ai-interview-with-ricardo-baeza-yates/

About the Author

Ricardo Baeza-Yates is Director of Research at the Institute for Experiential AI of Northeastern University. He is also a part-time Professor at Universitat Pompeu Fabra in Barcelona and Universidad de Chile in Santiago. Before, he was the CTO of NTENT, a semantic search technology company based in California, and prior to these roles, he was VP of Research at Yahoo Labs, based in Barcelona, Spain, and later in Sunnyvale, California, from 2006 to 2016. He is co-author of the best-selling textbook *Modern Information Retrieval* published by Addison-Wesley in 1999 and 2011 (2nd edn), which won the ASIST 2012 Book of the Year award. From 2002 to 2004 he was elected to the Board of Governors of the IEEE Computer Society and between 2012 and 2016 was elected to the ACM Council. Since 2010 he has been a founding member of the Chilean Academy of Engineering. In 2009 he was named ACM Fellow and in 2011 IEEE Fellow, among other awards and distinctions. He obtained a PhD in computer science from the University of Waterloo, Canada, in 1989, and his areas of expertise are web search and data mining, information retrieval, bias and ethics on AI, data science and algorithms in general. Regarding responsible AI, he is actively involved as an expert in many initiatives, committees, and advisory boards all around the world: Global Partnership on AI, Global AI Ethics Consortium, ACM's US Technology Policy Committee, and IEEE's Ethics Committee. He is also a co-founder of OptIA in Chile, an NGO devoted to algorithmic transparency and inclusion, and a member of the editorial committee of the new Springer journal *AI and Ethics*, in which he co-authored an article highlighting the importance of research freedom on AI ethics.

https://doi.org/10.1017/S1062798723000145 Published online by Cambridge University Press

Digital Humanism Initiative
Technology for Humanity

SPEAKER

Second African Digital Humanism Summer School, Accra, Ghana, 31 August – 05 September 2026

Victor de Boer

Victor de Boer is Associate Professor (UHD) at the Vrije Universiteit Amsterdam (VU), where he co-leads the User-Centric Data Science group of the Computer Science department. He is also a co-director of the Cultural AI Lab. In his research, he combines (Semantic) Web technologies with Human-Computer Interaction, Knowledge Representation and Information Extraction to tackle research challenges in various domains. These include Cultural Heritage, Digital Humanities and ICT for Development (ICT4D).

Lecture 10: *AI Knowledge Graphs for Decolonizing Cultural Heritage*

Reader article:

Marieke van Erp and Victor de Boer: A Polyvocal and Contextualised Semantic Web.
In R. Verborgh, K. Hose, H. Paulheim, P.-A. Champin, M. Maleshkova, O. Corcho, P. Ristoski, M. Alam (Eds.): Proceedings ESWC 2021 (18th International Conference on The Semantic Web), LNCS 12731, pp. 506–512. Springer Nature, Cham, Switzerland, 2021. https://doi.org/10.1007/978-3-030-77385-4_30

A Polyvocal and Contextualised Semantic Web

Marieke van Erp[1(✉)] and Victor de Boer[2]

[1] KNAW Humanities Cluster, Amsterdam, The Netherlands
marieke.van.erp@dh.huc.knaw.nl
[2] Vrije Universiteit Amsterdam, Amsterdam, The Netherlands
v.de.boer@vu.nl
https://mariekevanerp.com

Abstract. Current AI technologies and data representations often reflect the popular or majority vote. This is an inherent artefact of the frequency bias of many statistical analysis methods that are used to create for example knowledge graphs, resulting in simplified representations of the world in which diverse perspectives are underrepresented. With the use of AI-infused tools ever increasing, as well as the diverse audiences using these tools, this bias needs to be addressed in both the algorithms analysing data, as well as in the resulting representations. In this *problems to solve before you die* submission, we explain the implications of the lack of polyvocality and contextual knowledge in the semantic web. We identify three challenges for the Semantic Web community on dealing with various voices and perspectives as well as our vision for addressing it.

Keywords: Culturally aware AI · Polyvocality · Contextualisation · Bias

1 Introduction

Biases in data can be both explicit and implicit. Explicitly, 'The Dutch Seventeenth Century' and 'The Dutch Golden Age' are pseudo-synonymous and refer to a particular era of Dutch history. Implicitly, the 'Golden Age' moniker is contested due to the fact that the geopolitical and economic expansion came at a great cost, such as the slave trade. A simple two-word phrase can carry strong contestations, and entire research fields, such as post-colonial studies, are devoted to them. However, these sometimes subtle (and sometimes not so subtle) differences in voice are as yet not often found in knowledge graphs.

One of the reasons is that much of the knowledge found in knowledge graphs is mined automatically and current AI technologies (and their ensuing data representations) often reflect the popular or majority vote. This is an inherent artefact of the frequency bias of many statistical analysis methods that are used to create for example knowledge graphs, resulting in simplified representations

Supported by cultural-ai.nl.

R. Verborgh et al. (Eds.): ESWC 2021, LNCS 12731, pp. 506–512, 2021.
https://doi.org/10.1007/978-3-030-77385-4_30

of the world in which diverse perspectives are underrepresented. With the use of AI-infused tools ever increasing, as well as the diverse audiences using these tools, this bias needs to be addressed in both the algorithms analysing data, as well as in the resulting representations.

Conversations around data bias and polyvocality are taking place in for example the cultural heritage domain (cf. [14,15]) and computational linguistics communities (cf. [8,11]), but do not yet seem mainstream in the Semantic Web discourse. In 2010, Hendler and Berners-Lee already recognised that current knowledge mining mechanisms can bias results and recommend 'Making the different ontological commitments of competing interpretations explicit, and linked together', as this 'can permit different views of data to be simultaneously developed and explored.' [9]. Veltman noted in her 2004 paper that '[Dynamic knowledge will] allow us to trace changes of interpretation over time, have new insights and help us to discover new patterns in knowledge' [20]. Context was also flagged as the next frontier in knowledge representation at the 'Knowledge graphs: New directions for knowledge representation on the semantic web' Dagstuhl seminar in 2018 [5]. While it is known that large knowledge graphs are not always balanced (cf. [12]), that links between linked data resources contain biases (cf. [19]), and that research based on such resources might favour a Western perspective (cf. [6]), the creation of contextualised and polyvocal knowledge graphs has only gained modest traction.

As data-driven applications are permeating everyday life, it is necessary that these applications can serve as many and as diverse audiences as possible. A Dutch visitor to the Rijksmuseum in Amsterdam could recognise the Christian context of a painting depicting Saint Christopher even if s/he does not know that he is the patron saint of travellers. For a Japanese visitor, this might not be immediately apparent, and s/he needs more context to connect this to her/his cultural context, to for example recognise the parallels to Jizō, protector of travellers in some Buddhist traditions. For most other applications (e.g. in the domains of food, health, policy and transportation) a greater awareness of cultural and contextual frameworks of users would be crucial to increase engagement and understanding. Thus far, one of the ways machine learning driven approaches have dealt with this, is to add more data, at the risk of flattening nuances. At times, this has led to painful situations, for example when a Twitter trained chatbot started publishing racist, sexist and anti-semitic Tweets within a day of being released [23]. We argue that addressing this by more intelligently organised datasets to identify bias, and multiple perpectives and contexts is needed. Many of the puzzle pieces for a polyvocal and contextualised Semantic Web are already in place. The different local Wikipedias for example present different perspectives [4,24] with varying levels of in- and between group biases [1] but this is often still expressed implicitly.

In this *problems to solve before you die* submission, we explain the implications of the lack of polyvocality and contextual knowledge in the semantic web, as well as our vision for addressing it. The examples we use come from cultural heritage datasets as we work with those and some of the bias issues are amplified

in there due to their longue durée,[1] but these issues and our proposed approach also apply to other domains.

2 What Is a Voice?

Data is not objective, but rather created from a particular perspective or view, representing a *voice*. These perspectives can be informed by cultural, historical or social conventions or a combination of those. Often, these different perspectives are contiguous, rather than disjoint. A cultural view on for example the 'The Dutch Golden Age' is that it was an era in which the Dutch economy and scientific advancements were among the most acclaimed in the world, laying the foundations for the first global multinational corporation and shaping Dutch architecture. A historical view could be that it was an era of unbridled opportunity and wealth for many in the Dutch Republic. A social view could focus on the pride associated with the achievements of this era. However, none of these views tells the entire story. The Dutch Republic maintained several colonies in Africa and Asia and was heavily involved in slave trading; the other side of the coin of wealth and pride.

These different views often come to us through objects in cultural heritage collections which can often be different things throughout their life. For example, many cult figures and other objects were taken and brought back home by missionaries -in some cases to learn about 'the other' from a euro-centric view. Such items were often (dis)qualified as "fetishes" or "idols" and Christian converts were expected to stay away converted from them. Later, these objects were removed from their original ritual settings and became part of ethnographic collections in missionary exhibitions or were sold as works of art. These reframings represent new perspectives on the same object and depend on time, culture and the object provenance. Data models that are perspective-aware should be able to trace the various changes that objects undergo in their trajectory from their original uses in indigenous religious practices into museum collections [13].

A polyvocal Semantic Web provides opportunities, models and tools to identify, represent and show users different perspectives on an event, organisation, opinion or object. Furthermore, identified individual perspectives need to be connected and clustered to be lifted from an individual perspective to a (representative) voice of a group. Identifying, representing and using such groups with fuzzy boundaries that change over time is one of the core challenges.

3 How Can the Semantic Web Deal with Multiple Perspectives and Interpretations?

We identify three main challenges to support polyvocality on the Semantic Web: identification, representation and usage. We describe these challenges below, along with the current state of the art and promising research directions for addressing them.

[1] Some collections we work with have been gathered over a span of over 200 years, and the objects they describe go back even longer.

3.1 Identifying Polyvocality in Data Sources

Data sources of various modalities can represent a singular 'voice', and provide a specific perspective on the world. However, within and across data sources, there can also be multiple voices present. Moreover, when aggregating datasets, such voices can become lost in the process. The challenge here is to identify such voices. For natural language data sources, as well as for methods that extract information from structured sources, this requires Information Extraction methods that are 'bias' or 'voice'-aware. In cases where datasets are the result of aggregations, such methods should be able to (re-)identify the separate voices in the combined source.

In cases where human data providers are involved, the voices of the individual users should be maintained. These persons can also represent cultural or societal groups and this information should be available for subsequent representations. As an example, consider a crowdsourcing effort to annotate a Polynesian object in a European museum. Here, annotations provided by the European public on the one hand and annotations provided by members of the source community on the other hand will represent various perspectives. Methods that aggregate such annotations should retain these voices. An example of a method that does this for individual annotators is CrowdTruth [2], which maintains so-called 'disagreement' between annotators. Such methods can be expanded to retain this disagreement to the extent to which it represents a 'voice'.

What holds for these non-professional annotators, also holds for professionals involved in metadating or interpreting information. Different perspectives provided by such professionals should also be maintained in the information processing pipeline. Again, in the cultural heritage domain, with movements towards a diversification of the museum professionals gaining traction [18,22], the different perspectives that such professionals provide should be maintained in the object metadata.

3.2 Representation of Polyvocality: Datamodels and Formalisms

Identifying, extracting and retaining polyvocal information in data sources is one aspect, but this is meaningless if the various voices cannot be represented in the data structures used. Luckily, the Knowledge Graph as a data model and the Semantic Web as an information architecture provide excellent opportunities for maintaining various viewpoints on one subject. Its network structure and distributed nature is well-equipped to deal with such viewpoints. One example is found in Europeana, where the Europeana Data Model [10] allows for multiple publishers to provide information about one cultural heritage object. Building on the OAI-ORE model[2], each data provider provides a "Proxy" resource that represents that object in the context of that provider and all metadata is attached to that Proxy resource, rather than the object representation itself. This allows to have multiple perspectives on a single real-world object. This model has

[2] http://www.openarchives.org/ore/1.0/datamodel.

also been applied to represent multiple and potentially disagreeing biographical descriptions of persons [16]. In the NewsReader project,[3] information on events was mined from newspaper articles. Multiple reports on the same event were identified via in- and cross-document coreference resolution and presented using named graphs to group statements about a particular event. Provenance information is then attached to these named graphs to represent the viewpoints of various providers [7,17].

Where a voice corresponds to the view of a data publishers, such models appear adequate. At the other end of the spectrum, we see provenance at the level of the individual annotator or even the individual statement. WikiData for example, represents each statement as a tuple which includes the metainformation about that statement [21]. This allows for the recording of very precise provenance of statements, to the individual level.

The view of a data provider and the view of the individual annotator represent two points on the spectrum that can inform how to represent the more collective and elusive social-cultural voice. Here, we need models and design patterns on how voices can be represented. This means that not only "what is expressed" should be represented, but also "which worldview does this represent". One direction is that of "data lenses" as deployed for example in the OpenPhacts project [3]. Here sets of statements are annotated with provenance information, which in turn represent the scientific worldview of different types of end-users. In our case, we would like such lenses to correspond to various voices regardless of the end-user.

3.3 Usage of Polyvocal Knowledge

Representing and maintaining polyvocal knowledge is meaningless if it cannot be used in end-user facing applications. With the voices being represented in the knowledge graphs, different viewpoints are available for such applications. How to provide access to this more complex information in a meaningful way is a third major research challenge. The Europeana, Wikidata and GRaSP data models are already complex to query, even for Semantic Web experts. Adding another layer, that of the voice or lenses, potentially further complicates this.

Personalisation can be supported by these different views. Developers of end-user applications should take care that such personalisation does not lead to 'filter bubbles'. Effective communication of the various views that exist on specific resources is crucial. One example is objects that are assigned to different categories depending on the cultural context. Many objects in ethnography museums are currently categorised based on older (colonial) classification schemes, that might not represent currently held views, and moreover, such views are likely to differ per culture. If such polyvocal information about the object is maintained, it can be shown to the users in digital interfaces as well as in physical locations.

What is needed are design guidelines and patterns for visualisation of polyvocal Knowledge as well as reusable tools and methods.

[3] https://newsreader-project.eu.

4 Discussion

Creating a polyvocal and contextualised Semantic Web will be a community effort driven by interdisciplinary teams. Whilst the Cultural AI consortium is embedded in different cultural heritage institutions in the Netherlands that represent different voices, we have to be aware of the danger that this remains a singular (Western/Northern) voice. To create globally relevant and culturally aware semantic web resources, diversity and inclusivity is key, and for this we can for example take cues from Black in AI[4] and Widening NLP.[5] The Semantic Web community is a diverse one, as our conference attendants, organising and programme committees, editorial boards and mailing lists show. Different voices are present, so let us make our resources, models and projects reflect that.

Acknowledgements. This paper builds upon the core concepts and research agenda of the Cultural AI Lab (https://www.cultural-ai.nl/), the authors would like to thank the co-creators of this research agenda: Antal van den Bosch, Laura Hollink, Martijn Kleppe, Johan Oomen, Jacco van Ossenbruggen, Stephan Raaijmakers, Saskia Scheltjens, Rosemarie van der Veen-Oei, and Lotte Wilms. We also want to thank our colleagues of the "Culturally Aware AI" project funded by the Dutch Research Council's Responsible use of AI programme, from the SABIO project, funded by the Dutch Digital Heritage Network, and from "Pressing Matter: Ownership, Value and the Question of Colonial Heritage in Museums", funded through the National Science Agenda of the Dutch Research Council.

References

1. Álvarez, G., Oeberst, A., Cress, U., Ferrari, L.: Linguistic evidence of in-group bias in English and Spanish Wikipedia articles about international conflicts. Discourse Context Media **35**, 100391 (2020)
2. Aroyo, L., Welty, C.: Truth is a lie: crowd truth and the seven myths of human annotation. AI Mag. **36**(1), 15–24 (2015)
3. Batchelor, C., et al.: Scientific lenses to support multiple views over linked Chemistry data. In: Mika, P., et al. (eds.) ISWC 2014. LNCS, vol. 8796, pp. 98–113. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-11964-9_7
4. Beytía, P.: The positioning matters: estimating geographical bias in the multilingual record of biographies on Wikipedia. In: Companion Proceedings of the Web Conference 2020, pp. 806–810 (2020)
5. Bonatti, P.A., Decker, S., Polleres, A., Presutti, V.: Knowledge graphs: new directions for knowledge representation on the semantic web (dagstuhl seminar 18371). In: Dagstuhl Reports, vol. 8. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik (2019)
6. Farda-Sarbas, M., Mueller-Birn, C.: Wikidata from a research perspective-a systematic mapping study of wikidata. arXiv preprint arXiv:1908.11153 (2019)

[4] https://blackinai2020.vercel.app/.
[5] http://www.winlp.org/.

7. Fokkens, A., Vossen, P., Rospocher, M., Hoekstra, R., van Hage, W.R.: GRaSP: grounded representation and source perspective. In: Proceedings of the Workshop Knowledge Resources for the Socio-Economic Sciences and Humanities associated with RANLP 2017, pp. 19–25. INCOMA Inc., Varna, September 2017. https://doi.org/10.26615/978-954-452-040-3_003
8. Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J.W., Wallach, H., Daumé III, H., Crawford, K.: Datasheets for datasets. arXiv preprint arXiv:1803.09010 (2018)
9. Hendler, J., Berners-Lee, T.: From the semantic web to social machines: a research challenge for AI on the world wide web. Artif. Intell. **174**(2), 156–161 (2010)
10. Isaac, A.: Europeana data model primer (2013)
11. Jo, E.S., Gebru, T.: Lessons from archives: strategies for collecting sociocultural data in machine learning. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, pp. 306–316 (2020)
12. Konieczny, P., Klein, M.: Gender gap through time and space: a journey through Wikipedia biographies via the Wikidata human gender indicator. New Media Society **20**(12), 4608–4633 (2018)
13. Meyer, B.: Idolatry Beyond the Second Commandment: Conflicting Figurations and Sensations of the Unseen. Bloomsbury, London (2019)
14. Miles, R.S., Tzialli, A.O.E., Captain, E. (eds.): Inward Outward, Critical Archival Engagements with Sounds and Films of Coloniality - A Publication of the 2020 Inward Outward Symposium (2020). https://doi.org/10.18146/inout2020
15. National Museum for World Cultures: Words matter: An unfinished guide to word choice in the cultural sector. https://www.tropenmuseum.nl/nl/over-het-tropenmuseum/words-matter-publicatie (2018)
16. Ockeloen, N., et al.: BiographyNet: managing provenance at multiple levels and from different perspectives. In: LISC@ ISWC, pp. 59–71. Citeseer (2013)
17. Rospocher, M., et al.: Building event-centric knowledge graphs from news. Journal of Web Semantics **37**, 132–151 (2016)
18. Smith, C.E.: Decolonising the museum: The national museum of the American Indian in Washington, D.C. (2005)
19. Tiddi, I., d'Aquin, M., Motta, E.: Quantifying the bias in data links. In: Janowicz, K., Schlobach, S., Lambrix, P., Hyvönen, E. (eds.) EKAW 2014. LNCS (LNAI), vol. 8876, pp. 531–546. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-13704-9_40
20. Veltman, K.H.: Towards a semantic web for culture. J. Digital Inf. **4**(4) (2004)
21. Vrandečić, D., Krötzsch, M.: Wikidata: a free collaborative knowledgebase. Commun. ACM **57**(10), 78–85 (2014)
22. Wei, L.: Crossing the ethnic boundaries: diversity promotion effort of museum professionals in the United Kingdom. In: Proceedings of the Papers, CAAS 7th Conference at TUFS, October 2016
23. Wolf, M.J., Miller, K.W., Grodzinsky, F.S.: Why we should have seen that coming: comments on Microsoft's Tay "Experiment", and wider implications. ORBIT J. **1**(2), 1–12 (2017)
24. Zhou, Y., Demidova, E., Cristea, A.I.: Who likes me more? Analysing entity-centric language-specific bias in multilingual Wikipedia. In: Proceedings of the 31st Annual ACM Symposium on Applied Computing, pp. 750–757 (2016)

SPEAKER

Second African Digital Humanism Summer School, Accra, Ghana, 31 August – 05 September 2026

Prosper Simbarashe Maguchu

Prosper Simbarashe Maguchu is Senior Lecturer and Researcher in Public International Law, at the Law Faculty of Vrije Universiteit Amsterdam, Netherlands. He is also a Senior Advisor on International Development at the VU Center for International Cooperation, in Amsterdam. Previously he worked as an international coordinator for the Anti-Corruption Act Southern Africa and as a senior projects lawyer with the Zimbabwe Human Rights Forum specializing in public interest litigation on behalf of torture and organized violence victims. His main research interest are in the area of financial crimes and human rights.

Lecture 11: *A Human Rights Approach to Fight Corruption – The Role of Ubuntu*

Reader article:

Prosper Simbarashe Maguchu: Localizing the Human Rights-Based Approach to Fight Corruption: The Role of Ubuntu. Cross-cultural Human Rights Review Vol. 2 (Issue 1), pp 5-22, September 2020. https://doi.org/10.52854/cchrr.37

Cross-cultural Human Rights Review (2020) Vol.2:1 pp. 5-22
ISSN: 2666-3678 (Online) 2666-366X (Print)
Review homepage: https://www.cchrreview.org

LOCALIZING THE HUMAN RIGHTS-BASED APPROACH TO FIGHT CORRUPTION: THE ROLE OF UBUNTU

PROSPER MAGUCHU *

ABSTRACT

Claims by many experts on the connection between corruption and human rights, and especially the realization that corruption undermines the enjoyment of human rights, have led practitioners to advocate a human rights-based approach to corruption. However, it comes at a time where the global human rights movement is under assault, this contribution addresses the emerging localization discourse in human rights. Researchers and campaigners are adapting the international human rights system to local institutions and meanings in a process of "vernacularization". This is by taking the needs of the community and the language that makes sense locally as the entry point of human rights advocacy. The question that arises is what role can the local understanding of human rights play? This contribution suggests answers to this question by using the African concept of ubuntu (humanness) to reinforce measures against corruption.

KEYWORDS

Africa, Corruption, Human Rights-Based Approach, Localization, *Ubuntu*

* Dr Prosper Maguchu is a Project Manager for Justice and the Rule of Law Projects at the Center for International Cooperation. He is also an Assistant Professor at the Faculty of Law, Vrije Universiteit Amsterdam, The Netherlands.
Email: p.s.maguchu@vu.nl

** The author expresses his thanks to the anonymous reviewer for the very lucid insights, Professor Wouter Werner (Vrije Universiteit Amsterdam) and Nyashadzamwari Vera (Oslo Metropolitan University) for the exchange of thoughts and comments on parts of the draft text.

1. INTRODUCTION

Over the past decade, human rights activists and anti-corruption campaigners have been occupied with the connection between corruption and human rights. Albiny-Lackey is correct when he points out that practitioners from both the field of human rights and anti-corruption "have grown increasingly comfortable on one another's turf, and the lines that divide their work have become quite porous."[1] This is because, "human rights activists are learning to grapple with the human rights impacts of corruption, and anti-corruption campaigners frequently deploy a human rights analysis to bolster their case for reform."[2]

On the international plane, in 2019 the UN High Commissioner for Human Rights (OHCHR) in collaboration with the Geneva based Centre for Civil and Political Rights developed a *Practitioners Guide on Human Rights and Countering Corruption* that "provides guidance as well as practical recommendations on effectively integrating human rights into anti-corruption efforts."[3]

Similarly, international non-governmental organizations such as Transparency International (TI), the world's largest anti-corruption watchdog, have been leading the campaign to mainstream anti-corruption into human rights work.[4] In a 2009 joint publication with the now defunct International Council on Human Rights Policy (ICHRP), TI issued a thorough publication that focuses on "where and how the use of a human rights framework might strengthen national and local anti-corruption programmes and at how key human rights principles can be operationalized in anti-corruption work."[5]

What is more, legally binding international instruments are also reflecting the human rights-based approach in their texts. For instance, consider the 2003 United

[1] Chris Albin-Lackey, 'Corruption, Human Rights, and Activism: Useful Connections and Their Limits' in Dustin Sharp (ed) *Justice and Economic Violence in Transition* (Springer 2014) 139.

[2] Ibid 139.

[3] Divya Prasad and Làzarie Eeckeloo, *Corruption and Human Rights Integrating Human Rights in the Anti-Corruption Agenda: Challenges, Possibilities and Opportunities* (Centre for Civil and Political Rights 2019).

[4] See TI website on <https://www.transparency.org/en/> [last accessed 24 September 2020].

[5] International Council on Human Rights Policy, *Corruption and Human Rights: Making the Connection* (International Human Rights Council 2009).

Nations Convention Against Corruption (UNCAC). Although the UNCAC does not explicitly mention the words "human rights" in its text there are allusions to some rights language in the preamble and articles.[6] By way of illustration, the first paragraph of the UNCAC Preamble expresses concern about the seriousness of problems and threats posed by corruption to the stability and security of communities, by undermining the institutions and values of democracy, ethical values and justice and jeopardizing sustainable development and the rule of law.[7] Similarly, UNCAC's Article 11 refers to the independence of the judiciary and Article 13 calls for "respecting, promoting and protecting the freedom to seek, receive, publish and disseminate information concerning corruption...".[8] These provisions impliedly protect fundamental rights recognized by international treaties and domestic constitutions of many states such as the right to justice,[9] right to development,[10] right to information,[11] right to judicial independence,[12] and freedom of association.[13]

Academics claim mainstreaming corruption in human rights can assist to humanize anti-corruption. For instance, longitudinal studies by de Castro e Silva, comparing the rationality and the values addressed in both corrupt and non-corrupt cases revealed that, including a human rights-based approach can reinforce measures

[6] UN Coalition Against Corruption, 'A human Rights-Based Approach to Corruption' Document submitted by the UNCAC Coalition, a non-governmental organization not in consultative status with the Economic and Social Council on 31 August 2018 < https://uncaccoalition.org/wp-content/uploads/HR-based-approach.pdf> [last accessed 28 September 2020] 2.

[7] UN Coalition Against Corruption 2003 Preamble.

[8] UN Coalition Against Corruption (n 6) 2.

[9] Article 7 of the African Charter on Human and Peoples' Rights 1981 (African Charter); Article 6 of the European Convention for the Protection of Human Rights and Fundamental Freedoms (ECHR) 1950; Article 14 of the International Covenant on Civil and Political Rights 1966.

[10] Article 22 of the African Charter (n.9), See also the UN Declaration on the Right to Development 1986.

[11] Article 9 of the African Charter (n.9); Article 10 of the ECHR (n.9).

[12] Article African Charter on Human and Peoples' Rights, Article 47 European Charter on Human Rights.

[13] Article 10 African Charter on Human and Peoples' Rights, Article 12 European Charter on Human Rights.

against corruption.[14] Such approach (i) enhances the justiciability of economic and social rights; (ii) it is a radical shift from the insufficient criminal approach to a focus on the social harm; and (iii) it is a more satisfactory approach to the far-reaching effects of corruption and inequality.[15] De Castro e Silva believes by treating anti-corruption and human rights as mutually reinforcing, practitioners can create a 'legal empowerment strategy', that focuses particularly on the social accountability dimension. It strengthens the disadvantaged, and fights the encroachment caused by corruption on the enjoyment of human rights, especially economic and social rights.[16]

The inclusion is not without opposition. Peters has summarized some of the critical accounts:

> Some international lawyers might complain that this smacks of 'human rightism' or of a 'hubris' of international human rights. Indeed, there is a risk of overusing human rights language. Therefore, the human rights-based approach to corruption should not be employed as a panacea. The language of law generally (and of rights, more particularly) is a limited one, as the critique of the human rights-based approach to corruption points out.[17]

Moreover, other commentators like Pellet have argued that human rights law "is, and can only be, the art of the possible, and by wanting to ask the impossible of it, the 'human rightists' harm the cause that they intend to defend more than they serve it."[18] Pellet further describes the processes of incorporating human rights approaches as a dangerous tendency of human rightism, which he compares to "wishful thinking."[19] Similarly, Rose emphasizes that, applying a human rights lens to combating corruption might not be the best approach, "because human rights law has its

[14] Bruna de Castro e Silva, 'Humanizing (Anti)Corruption: The Socio-Legal Values of a Human Rights-Based Approach to Corruption' [2019] Kyiv-Mohyla Law and Politics Journal 59, 60.

[15] ibid 62-64.

[16] Ibid 67.

[17] Anne Peters, 'Corruption as a Violation of International Human Rights' (2018) 29 European Journal of International Law 1251, 1286.

[18] Alain Pellet, '"Human Rightism" and International Law' (2000) 10 The Italian Yearbook of International Law 1, 5.

[19] ibid.

limitations as a vocabulary describing the harms of corruption, there is potential for lawyers and activists to overuse or misappropriate […] this rhetoric."[20]

Aside from these criticisms, there is indeed, an unambiguous relationship between corruption and human rights. To this end, this contribution will not focus so much on illuminating the link between the two as this has been thoroughly discussed elsewhere.[21] However, it will discuss the potential of localizing human rights - one of the emerging discussions in the field of human rights - in particular within the concept of *ubuntu* and how it can be used in context of anti-corruption.

2. CORRUPTION AND THE HUMAN RIGHTS-BASED APPROACH

How does the human rights-based approach to anti-corruption work? By and large, the human rights-based approach and the criminal justice-based approach complement each other because they emphasize different harms and responsibilities. Below are some of the oft cited benefits of a human rights-based approach to corruption.

First, the human rights-based approach advocates recognition of corruption as a human rights concern. The implication of this is that it complements existing anti-corruption efforts by enabling international and regional human rights courts, as well as domestic courts, to hear cases concerning the rights of individuals that have been violated by corrupt acts.[22]

In addition to international and regional human rights courts, there are several other relevant mechanisms and processes in the human rights system, such as special rapporteurs, national agencies and non-governmental organizations that can be harnessed to fight corruption. Davis explains how these other different mechanisms have varying capacity, expertise and claims to legitimacy, but, collectively, they can

[20] Cecily Rose, 'The Limitations of a Human Rights Approach to Corruption' (2016) 65 International and Comparative Law Quarterly 405, 438.

[21] International Council on Human Rights Policy (n 5); Martina Boersma, *Corruption: A Violation of Human Rights and a Crime under International Law?* (Intersentia 2012); Kolawole Olaniyan, *Corruption and Human Rights Law in Africa* (Hart Publishing 2014); Peters (n 17).

[22] Peters (n 17) 1278; Gillian Dell, 'Addressing Impunity for Grand Corruption: What Are the Options?' Transparency International, 04 March 2020, <https://www.transparency.org/en/blog/addressing-impunity-for-grand-corruption-what-are-the-options#> [last accessed 24 September 2020].

bring impressive resources to bear.[23] For instance, Davis is of the view that some mechanisms that have fact-finding powers and capacities can help to gather information about corruption.[24] Alternatively, other processes can be employed to engage with the regular monitoring of particular public entities or topics.[25] Moreover, human rights actors can play a useful role in checking information gathered from other sources.

Second, the human rights-based approach moves the victim from the periphery to the centre. Whereas the criminal justice based-approach is centered on the State and the public it represents, on the other hand, a human rights-based approach to corruption means adding a perspective that moves the victim to the centre of the fight against corruption. It does so by highlighting the negative impacts that corruption has on an individual; on groups of individuals typically affected by corruption such as the most vulnerable or marginalized groups; and on society overall.[26]

Thus, addressing corruption through a human rights-based approach as envisioned by Rothstein and Varraich, pays special attention to the victims and the human cost that is at stake.[27] Such a holistic approach focuses not only on the economic and criminal consequences of corruption, but also on the victims, especially those belonging to vulnerable or marginalized groups. It also helps to empower the individuals affected by corruption and to transform them from being bystanders into actors in the fight against the scourge.[28] In this sense, a human rights-based approach

[23] Kevin E Davis, 'Corruption as a Violation of International Human Rights: A Reply to Anne Peters' (2019) 29 European Journal of International Law 1289–1296.

[24] ibid 1295.

[25] Ibid.

[26] UNHRC, Final Report of the Human Rights Council Advisory Committee on the Issue of the Negative Impact of Corruption on the Enjoyment of Human Rights, A/HRC/AC/12/CRP.3.

[27] Bo Rothstein and Aiysha Varraich, *Making Sense of Corruption* (Cambridge University Press 2017) 60.

[28] United Nations Development Programme Oslo Governance Centre 'The Impact of Corruption on the Human Rights Based Approach to Development' September 2004, <http://www.albacharia.ma/xmlui/bitstream/handle/123456789/30538/0284The_Impact_of_Corruption_on_the_Human_Rights_Approach_to_Development(2005)r.pdf?sequence=1> [last accessed 23 September 2020].

to corruption lays bare the cumulative damages of corruption such as oppression and inequality.[29]

Put differently, if done carefully, bringing claims and cases related to corruption before international and regional human rights processes and mechanisms could help to address corruption by providing adequate remedies for the victims.[30] Spalding has written extensively on similar arguments, such as focusing on the need for restorative justice and victim reparation approaches to anti-corruption efforts in the private sector.[31]

Third, and more importantly, the human rights-based approach puts emphasis on how the State has breached its obligations under international human rights law towards the public by failing to protect it from corruption.[32] The aforementioned groundbreaking research by TI and ICHRP sets out an operational framework that attempts to establish when a corrupt act leads to a violation of human rights.[33] In other words, the framework provides a technique for analysing corruption in human rights terms.[34]

According to the ICHRP corruption-related human rights violations can be summarized as follows:

(1) Identify the corrupt practice and establish what corrupt act is involved (bribery, embezzlement, etc.).

(2) Identify perpetrator(s): state actor or someone acting in partnership with a government official (e.g. if a private party commits the violation).

[29] Peters (n 17) 1252.

[30] Divya Prasad and Làzarie Eeckeloo (n 3) 9.

[31] Andrew B Spalding, 'Restorative Justice for Multinational Corporations' (2015) 76 Ohio State Law Journal 357-407.

[32] International Council on Human Rights Policy (n 5); Angela Barkhouse, Hugo Hoyland and Marc Limon, *Corruption and Human Rights* (Universal Rights Group and Kroll 2018); Julio Bacio-Terracino 'Linking Corruption and Human Rights' (2010) 104 Proceedings of the Annual Meeting (American Society of International Law 243-246.

[33] International Council on Human Rights Policy (n 5) 24-29; See also Angela Barkhouse, Hugo Hoyland and Marc Limon (n 32) 4-5.

[34] International Council on Human Rights Policy (n 5) 24.

(3) Identify the state's human rights obligations, study the scope and content of the human right in question and establish what acts or omissions of the state were required by the right in question.
(4) Evaluate the causal link between the corrupt practice and the harm; establish how direct the connection is between the corrupt act and harm suffered by the victim on the one hand, and the content of the human right and the obligation required from the state on the other hand.
(5) Evaluate the responsibility of the state for the damage caused; and determine if the state had undertaken an effective investigation and prosecuted those found responsible; further determine which forms of reparation would be adequate for the given case (e.g. restitution, compensation, satisfaction, etc.). [35]

This framework provides a helpful and detailed clarification on how the human rights-based approach works in relation to anti-corruption. It demonstrates the normative potential for human rights to assist in the fight against corruption. I will now analyse how such human rights-based approach can be localized through the concept of *ubuntu*.

3. LOCALIZING HUMAN RIGHTS: THE ROLE OF UBUNTU

Despite the efforts that were made since the adoption of the Universal Declaration of Human Rights (UDHR) in 1948, to apply human rights on a global or universal scale,[36] attention has turned to the adaptation of human rights to local conditions. The reasons are varied. De Feyter and Parmentier have summarized this in their influential book *The Local Relevance of Human Rights*, as follows:

> It is, however, unclear whether the global regime that was so painstakingly developed over the last few decades is of much practical use at the local level to people confronted with the abuse of power and/or inhumane living conditions. This question is

[35] International Council on Human Rights Policy (n 5) 28-29.

[36] Koen De Feyter and Stephan Parmentier, 'Introduction: Reconsidering human rights from below' in Koen De Feyter, Stephan Parmentier, Christiane Timmerman, and George Ulrich (eds), *The Local Relevance of Human Rights* (Cambridge University Press 2011) 1.

> particularly important in the current era of globalisation, when economic and political institutions of different kinds shape and reshape the world at a rapid pace.[37]

Hence, globalization, among others has sparked an important impetus for the localization of human rights. Ako defined localization as:

> a process that examines how local communities engage with issues from a rights-based context, particularly from (but not limited to) deprived or disadvantaged positions, in order to seek relevance in a globalised context.[38]

Indeed, the discourse for the localization of human rights is not only coming from the hitherto deprived and disadvantages communities. The discourse has attracted its own specialized working papers series, *Localizing Human Rights Working Paper Series*.[39] This correctly identifies the importance of "the local relevance of human rights, particularly but not exclusively in non-Western contexts."[40] For instance, the experiences of localization are showing that human rights can make a difference even in England, "a country badgered by widespread skepticism towards anything international."[41]

In the same vain, African scholars have been working on the concept of *ubuntu* as an entry point for the localization of human rights in the continent. Whilst large numbers of empirical researches into *ubuntu* and human rights are present in South Africa, few and far between are seen in the rest of the continent. Although the African Charter on Human and Peoples Rights is being considered the main source of the

---

[37] ibid.

[38] Rhuks Ako, 'Enforcing Environmental Rights under Nigeria's 1999 Constitution: The Localisation of Human Rights in the Niger Delta Region' in Koen De Feyter, Stephan Parmentier, Christiane Timmerman, and George Ulrich (eds) *The Local Relevance of Human Rights* (Cambridge University Press 2011) 270-294.

[39] See Localising Human Rights Working Paper Series, <https://www.uantwerpen.be/en/research-groups/law-and-development/publications/localising-human-rights/> [last accessed 24 September 2020].

[40] ibid.

[41] Casla and Dalmeny, 'What Does this Mean Here? Localizing Human Rights in the UK' Open Global Rights, October 31, 2019 <https://www.openglobalrights.org/localizing-human-rights-in-the-UK/> [last accessed 25 September 2020].

concept of ubuntu, the South African constitution has become associated with the concept.[42] *Ubuntu* embraces all the notions of universal human interdependence, solidarity, and communalism which bolster collective survival. Values of *ubuntu* are known in other systems of jurisprudence although by different names.[43] As Granger observed:

> Though it is a South African philosophical terminology, the fundamental epistemological elements of Ubuntu are shared by all Africans communities, with some varying peculiarities due to space and time. Scholars (other scholars) have articulately identified [*Ubuntu*] as the concept [t]hat adequately encapsulates the African indigenous and traditional philosophy.[44]

The dominant script of *ubuntu* has been shaped by the Nguni languages (and cognate terms such as *botho* in *Sotho-Tswana* and *hunhu* in *Shona*).[45] There has been a steady stream of scholarship and popular writing that is promoting *Ubuntu* as a plausible local response to the global human rights standards.[46] Murithi argues that *ubuntu* advances a framework of human dignity or 'humanness' that resonates with the

[42] Marius Pieterse 'Traditional African Jurisprudence' in Christopher Roederer and Darrel Moellendorf (eds) *Jurisprudence* (Kluwer Academic Publishers 2004) 438, 438.

[43] Dial Dayana Ndima, 'Reconceiving African Jurisprudence in a Post-Imperial Society: The Role of *Ubuntu* in Constitutional Adjudication' (2015) 48 The Comparative and International Law Journal of Southern Africa 359, 366.

[44] Peter Genger, 'Combating Corruption with African Restorative Justice Tradition: Suggested Steps for Nigeria' (2011) 11 African Journal of Criminology and Justice Studies 20, 24.

[45] Thaddeus Metz, '*Ubuntu* as a Moral Theory and Human Rights in South Africa' (2011) 11 African Human Rights Law Journal 532, 533.

[46] Tim Murithi, 'A Local Response to the Global Human Rights Standard: The *Ubuntu* Perspective on Human Dignity' (2007) 5 Globalisation, Societies and Education, 277, 281; Nokulunga Queeneth Mkabela, 'Ubuntu as an Axiological Framework for Human Rights Education' (2014) 13 Journal of Indigenous Knowledge Systems 283 -291; Yusef Waghid, 'African Philosophy of Education as a Response to Human Rights Violations: Cultivating *Ubuntu* as a Virtue in Religious Education' (2014) 27 Journal for the Study of Religion 267–282; In contrast see Antony Oyowe, 'Strange Bedfellows: Rethinking Ubuntu and Human Rights in South Africa' (2013) 5 African Human Rights Law Journal 103-124.

notion of human rights.[47] Homonga has also argued that, "[a]rguably, concepts such as [*ubuntu*], which mitigates against western individualism, have the potential to enhance the cultural legitimacy of human rights and their enforceability in African societies."[48] In his advice to President Barak Obama, Campbell put forward that "a new concept of social collectivism [*Ubuntu*] must be the basis of economic, social, and political organisation if humans are to survive the challenges of the 21st century".[49] It is my attempt here to contribute to the body of scholarship that advocates for the inclusion of localized and cultural experiences in the concept of universal human rights. One of the major aims in this column piece is to consider the role *ubuntu* plays, or ought to play, in applying a human rights-based approach, in its interpretation and implementation against corruption.

It is therefore appropriate to define the concept of *ubuntu* here. Not everyone means the same thing by '*ubuntu*'. However, Muller et al have offered a succinct definition of *ubuntu* "as the African idea of personhood: persons depend on other persons in order to be".[50] This is summarized in the expression: *umuntu ngumuntu ngabantu*, that is to say, a person is a person through persons.[51]

That said, the concept of *ubuntu* remains one of the misunderstood topics in the African human rights discourse. For instance, its legal meaning is not settled. Indeed, as political commentator McKaiser observes, what:

[47] Murithi (n 46) 5.

[48] Chuma Himonga, 'The Right to Health in an African Cultural Context: The Role of Ubuntu in the Realization of the Right to Health with Special Reference to South Africa' (2013) 57 Journal of African Law 165, 165.

[49] Horace Campbell, *Barak Obama and 21st Century Politics: A Revolutionary Moment in the USA* (Pluto Press 2010) quoted in Horace Campbell, 'Human Rights, Livelihood and Ubuntu for the 21st Century' Pambazuka News, December 09, 2010 <https://www.pambazuka.org/pan-africanism/human-rights-livelihoods-and-ubuntu-21st-century> [last accessed 23 September 2020].

[50] Julian Muller, John Eliastan and Sheila Trahereds (eds) *Unfolding Narratives of Ubuntu in Southern Africa* (Routledge 2018) 1.

[51] ibid.

> [*ubuntu* means in a legal context] depends on what a judge had for breakfast, [and it is] a terribly opaque notion not fit as a normative moral principle that can guide our actions, let alone be a transparent and substantive basis for legal adjudication.[52]

Aside from the definitional problems, *ubuntu* has been further criticized even as a human rights concept. Among the criticisms brought forward are:

(1) its vagueness: although the word literally means humanness, it lacks the precision required in order to render a publicly-justifiable rationale for making a particular decision;
(2) its apparent collectivist orientation: with many suspecting that it requires some kind of group-think, uncompromising majoritarianism or extreme sacrifice for society, which is incompatible with the value of individual freedom that is among the most promising ideals in the liberal tradition; and
(3) its traditional origin: in that ideas associated with *ubuntu* grew out of smallscale, pastoral societies in the pre-colonial era whose world views were based on thickly spiritual notions such as relationships with ancestors (the 'living-dead').[53]

Despite this, Metz has convincingly defended the concept of *ubuntu* against some unsubstantiated, sometimes unscholarly attacks, through empirical research that shows how *ubuntu* can provide a "promising foundation for human rights".[54] To support this assertion it could be argued that, generally human beings find dignity by virtue of their capacity to belong to a community, being identified with and exhibiting solidarity with others.[55] The concept of ubuntu promotes solidarity through its values of "communality, respect, dignity, value, acceptance, sharing, co-responsibility, humaneness, social justice, fairness, personhood, morality, group solidarity, compassion, joy, love, fulfilment, conciliation, et cetera."[56] The recognition of the close

[52] Eusebius McKaiser, Public Morality: Is there Sense in Looking for Unique Definition of Ubuntu? Business Day 2 November 2009.

[53] Metz (n 45) 53.

[54] ibid.

[55] ibid, 60.

[56] Irma J. Kroeze, 'Doing Things with Values: The Case of Ubuntu' (2002) 2 Stellenbosch Law Review 252, 260.

link between the respect for human rights and human dignity count as a strong impetus that *ubuntu* can be a basis for the local strengthening of human rights. It thus seems reasonable to suggest that the concept of *ubuntu* in certain ways can be part of the 'localizing' solution to contemporary moral dilemmas in the fight against corruption.[57] The next section examines the use of *ubuntu* for the purposes of anti-corruption.

4. UBUNTU AND ANTI-CORRUPTION

The basis of using the concept of *ubuntu* as a stratagem against corruption is based on the understanding of the concept of *people first*. The concept of putting people first to fight corruption has been used in other jurisdictions. For instance, in Ireland the *Action Programme Putting People First* is a government policy for reform and development that puts a strong emphasis on accountability, as the bedrock of a properly functioning system of local democracy. It provides for better engagement with citizens and introduces new degrees of accountability, transparency and external scrutiny, as essential pillars of local democracy.[58] Likewise, in the US the recent anti-corruption legislation, "For the People Act" aims to give power back to the people as envisaged in the US Constitution.[59]

In South Africa, some experts familiar with *ubuntu* have convincingly argued that *ubuntu* can be used to fight corruption. Dumisa and Amao are of the view that using the concept of *ubuntu* to fight corruption "is achievable through the promotion of a stable and ethical leadership complemented with strong institutions and structures to detect and eliminate corruption."[60] Furthermore, they propose that, "the

[57] See Campbell (n 49).

[58] See Baile Atha Cliath et. al., *Putting People First - Action Programme for Effective Local Government* (Government of Ireland 2012) i.

[59] See Michael Sozan and Will Ragland, Recent Political Scandals the 'For the People Act' Would Prevent From Recurring' Center for American Progress, February 4 2019 <https://www.americanprogress.org/issues/democracy/news/2019/02/04/465792/recent-political-scandals-people-act-prevent-recurring/> [last accessed 21 September 2020].

[60] Siphesihle Dumisa and Olumuyiwa Babatunde Amao, 'The Utility of Moral Philosophy and Professional Ethics in the Fight against Corruption in South Africa: Any Role for Ubuntu?' (2015) 4 Journal of Conflict and Transformation 83, 105.

formation of a strong anti-corruption regulatory institution, premised on the philosophical foundation(s) of Ubuntu"[61] could help foster transparency and accountability in the public sector. It would do so by ensuring adherence to high moral standards such as common good, human dignity and mutual respect, all which militate against corrupt behavior. Coupled with the fact that the concept of *ubuntu* conveys amongst other attributes, 'humanness', 'good moral disposition' and 'a capacity of social self-sacrifice on behalf of others', this will arguably assist in putting people first as a method of fighting corruption.

However, more recently, cross-national research projects are attempting to isolate the direct correlation between the culture of *ubuntu* and corrupt practices in Africa.[62] In particular, factors found to be influencing corruption as a result of *ubuntu* have been explored by Van der Walt who argued that communalism as a cornerstone of *ubuntu*, can potentially make African societies prone to corruption,[63] for the following two reasons.[64]

First, communalism causes the distinction between private and public funds to become vague.[65] This is possibly due to the inherent philosophy in communalism that is against private property and supports collective ownership. Corruption becomes a matter of culture in general, in the sense that culture represents a set of informal institutions adopted by the population of a country. Corruption also becomes a matter of culture in particular, as a consequence of the cultural embodiment of the formal institution of "property rights".[66] In other words, because public authorities are left with discretionary powers over property ownership this may breed corruption.

[61] ibid.

[62] See for example Colm Allan, Roberts Mattes and Unathi Millie, 'Government Corruption Seen from the Inside: A Survey of Public Officials' Perceptions of Corruption in the Eastern Cape' (PSAM Research Paper 1 2002)

[63] B.J Van der Walt, *Understanding and rebuilding Africa* (ICCA 2003) 406.

[64] ibid.

[65] ibid.

[66] Davor Mance and Mario Pecaric, 'Relationship Between Property Rights Enforcement and Corruption' paper presented at the 16th International Scientific Conference on Economic and Social Development – The Legal Challenges of Modern World – Split, 1-2 September 2016, 777.

Therefore, clear and unambiguously defined property rights can help to reduce corruption.

Second, communalism puts people under enormous financial pressure.[67] Partly because of the concepts of loyalty and reciprocity, an individual is beholden to other members of the community. The far-reaching implication of this is that people are responsible to take care of their immediate and extended family.[68] This in turn leads to corruption.[69] Van der Walt's reveals that pressure exerted by the extended family system, might cause some people to engage in acts of corruption in order to meet their obligations. [70] The social moral code that underpins the doctrine of *ubuntu*, it has been argued, 'requires' of people to render 'services' and 'counter-services'.[71] This practice often gives rise to several typologies of corruption such as favoritism, nepotism, tribalism.[72]

In studies conducted in South Africa, participants at a workshop made surprising admissions that some senior officials are under pressure to appoint 'struggle-comrades' in jobs, regardless of their skills or qualifications.[73] This is in consonance with a study done in Benin, Niger and Senegal by Blundo and De Sardan which revealed that corruption is often tolerated, because it is legitimised by social practices such as favours, string-pulling, professional protectionism and nepotism, especially when family, neighbours or friends are involved.[74] Grobler and Joubert add that, although nepotism is unethical, African cabinets and public offices are often filled

[67] Van der Walt (n 63) 406.

[68] Petria M Theron, 'Corruption in Sub-Saharan Africa: A practical-theological response'(2013) 47 In die Skriflig/In Luce Verbi 1, 3.

[69] Van der Walt (n 63) 406.

[70] Maurice O. Dassah, 'Is there a Hole in the Bucket? Identifying Drivers of Public Sector Corruption, Effects and Instituting Effective Combative Measures' (2008) 43 Journal of Public Administration 37–62.

[71] Theron (n 68) 5.

[72] For a contrary view see in general Tokunboh Adeyemo, *Is Africa cursed?* (Christian Learning Materials Centre 1997); See also Wilbur O'Donovan *Biblical Christianity in modern Africa* (Paternoster Press 2000) 11; Mustafa Hussein, 'Combating Corruption in Malawi: An Assessment of the Enforcing Mechanisms' (2005) 14 African Security Review 91, 95.

[73] Theron (n 68) 7.

[74] Giorgio Blundo and Jean-Pierre Olivier De Sardan, *Everyday corruption and the state: Citizens and public officials in Africa* (David Philip Publishers 2006) 9, 12.

with members of the presidents' tribes to ensure the perpetuation of their power base.[75]

The results from these studies all seems to demonstrate the role of culture in ethnic ties and the possible influence it has on corruption. In particular, the social pressure that is placed on individuals as members of a community may force them to behave unethically. To give an illustration, a public official who shows favour to his own kith and kin may not consider this an act of corruption, nor as a deviation from rules, but as a family obligation. This is in line with studies revealing that corruption strongly correlates with "cultural variables", like family ties.[76]

Despite the criticisms levied against the role of culture and communalism within the concept of *ubuntu*, there is ample empirical research determined to demonstrate a neutral correlation between culture and corruption. These show that depending on the social and cultural context in which the transactions take place, "participants in these exchanges are rather nuanced about the activities they are involved in."[77] That is to say, the distinction between what is good and bad cultural practices is not always as straightforward or clear-cut.[78] Therefore, it seems plausible that "any definition of corruption should reflect cultural differences."[79] For instance, Kubiciel has warned against the imposition of "a one sided", Western definition of

[75] Elana Grobler and S Joubert S.J., 'Corruption in the Public Sector: The Elusive Crime' (2004) 17 Acta Criminologica 90, 94.

[76] Robert Klitgaard, 'On Culture and Corruption' paper was presented at the Public Integrity and Anti-Corruption workshop at Nuffield College, Oxford in June 2017 <https://www.bsg.ox.ac.uk/research/publications/culture-and-corruption> [last accessed 19 September 2020].

[77] Henk van de Bunt and Hans Nelsen, 'Corruption in Various Shapes and Sizes—Some Criminological Reflections' in Henk van de Bunt et. al. (eds) *International Law and the Fight Against Corruption* (Asser Press 2012) 10.

[78] ibid 10-11.

[79] Xandra E. Kramer, 'Private International Law Responses to Jurisdiction and Foreign Judgements and the International Fight against Corruption' in Henk van de Bunt et. al. (eds) *International Law and the Fight Against Corruption* (Asser Press 2012) 59.

corruption that seems to imply that citizens of less developed countries are more corrupt than citizens of more developed countries.[80]

Moreover, on the recognition of cultural roles in the determination of corruption, article 15 of the UNCAC, takes a flexible reference to 'undue advantage'.[81] Similarly, the explanatory report to the Council of Europe Criminal Law Convention "equally reflects the moral imperialist objection and excludes socially acceptable gifts from its coverage."[82] As Xander aptly puts,

> An undue advantage implies that the official would have acted differently in the absence of the advantage [and] is thus linked to an official's specific behavior. Cultural gift-giving practices, however, are deemed to be a matter of social curtesy: they do not necessarily constitute corruption as they are not aimed to alter a specific act of a public official. Hence, even the traditional definition of corruption could allow for the exclusion of social courtesy practices.[83]

Thus, understanding the cultural background of an action can impact whether it is considered corrupt. Arguably then, a tempered approach to *ubuntu* is necessary when employing it in the fight against corruption.

5. CONCLUSION

This contribution draws attention to the need to make human rights more locally relevant in the fight against corruption. If made more locally relevant, the human rights-based approach can be an effective tool to fight against the adverse effects of corruption on the enjoyment of human rights. In the context of Africa, the concept of *ubuntu* can be the most plausible way to localize human rights. The concept of *ubuntu* conveys among others the idea of 'humanness', 'good moral disposition' and 'a capacity of social self-sacrifice on behalf of others' which can help with putting

[80] Michael Kubiciel, 'Core Criminal Law Provisions in the United Nations Convention against Corruption'(2009) 9 International Criminal Law Review 139, 145.
[81] UN Coalition Against Corruption (n 6).
[82] Council of Europe, 'Explanatory Report to the Criminal Law Convention on Corruption' European Treaty Series No. 173, 8 para 9.
[83] Kramer (n 79) 60.

people first as a method of fighting corruption.[84] While the focus on culture and communalism within the concept of *ubuntu* is argued by some as fueling corruption, this column suggests that this should not always be the case. Actions have diverse meaning across different cultural contexts, which should be borne in mind when deciding if an action is corrupt.

[84] See Christian B.N. Gade, 'The Historical Development of the Written Discourses on Ubuntu' (2011) 30 South African Journal of Philosophy 307–308.

Digital Humanism Initiative
Technology for Humanity

SPEAKER

Second African Digital Humanism Summer School, Accra, Ghana, 31 August – 05 September 2026

Gaétan Hains

Gaétan Hains is a professor of computer science in the Faculty of Science and Technology at Université Paris-Est Créteil (UPEC) and a member of the Algorithms, Complexity, and Logic Laboratory (LACL). He is currently vice-president for digital affairs of the university. A former recipient of scholarships from Quebec, Canada, and the Commonwealth, he graduated from the University of Oxford and served as a faculty member and researcher at the University of Montreal and then the University of Orléans before joining UPEC in 2006. He has served as director of computer science laboratories, program manager for software programs at the French ANR, and head of R&D teams in the private sector. He is the author of more than 100 publications in Computer Science and supervised a dozen PhD students.

Lecture 12: *Low-Tech AI – Low Cost, Low Energy, High Social Value Applications*

Reader article:

Manpreet Kour, Gaetan Hains, and Mohamed Saber: Analysis of Bar Charts and Pie Charts for Blind or Visually Impaired Readers. Université Paris-Est Créteil (UPEC), Val de Marne, France, 2025. HAL Open Science Id: hal-05287303. Submitted 28 September 2025. https://hal.science/hal-05287303v1

Analysis of Bar Charts and Pie Charts for Blind or Visually Impaired Readers

Manpreet Kour, Gaetan Hains, Mohamed Saber

Analysis of Bar Charts and Pie Charts for Blind or Visually Impaired Readers

Manpreet Kour, Gaetan Hains, Mohamed Saber

LACL, Universite Paris Est Creteil (UPEC) and Orange Orléans
manpreet.kour@u-pec.fr,gaetan.hains@u-pec.fr,mohamed.saber@orange.com

Abstract. Data visualizations and analysis of diagrams such as bar charts and pie charts remain largely inaccessible to blind and visually impaired individuals, despite their widespread use in education, science, and public communication. In this work, we present a learning-based framework that automatically detects, extracts, and interprets the contents of bar charts, translating them into multi-modal outputs such as text, answers to queries and/or text-to-speech narratives Our system employs a YOLOv8 object detection model trained on a custom-generated dataset of annotated bar charts, capable of identifying key chart components, including bars,slices,angles axes, titles, and labels. Following detection, we integrate Optical Character Recognition (OCR) to decode textual labels and numerals, while a pixel-to-value mapping algorithm interprets bar heights relative to axis scales. The extracted information is dynamically translated into audio descriptions and interactive query-based speech outputs, enabling non-visual comprehension of chart data. To ensure usability, we designed the output flow to follow a hierarchical structure, from summary to specific values. Preliminary evaluations show that the system achieves high accuracy in detecting component parts of the chart and generating bar values, with promising feedback from blind users in pilot tests. This research contributes to inclusive AI, opening new directions for accessible data communication through the integration of computer vision, speech synthesis, and assistive reasoning. *abstract* environment.

Keywords: Bar chart accessibility; YOLO; object detection; optical character recognition (OCR); text-to-speech (TTS); sonification; blind and visually impaired users; assistive technology; AI for accessibility; document image understanding

1 Introduction

Bar charts are one of the foremost general shapes of visual communication for displaying categorical and quantitative information. They are widely used in academic writing, instructive content, trade reports, and open data campaigns. Despite their effortlessness and viability, bar charts remain fundamentally inaccessible to blind and visually impaired individuals. Screen readers are unfit to

translate inserted chart structures or extract significant numerical bits of knowledge from graphical components. As a result, blind and visually-impaired (BVI) users are often excluded from accessing data-rich content or restricted to depend on inactive elective content that is physically composed and regularly deficient or excessively rearranged.

Whereas there has been significant advance in making web substance open through screen reader, material design, and sound portrayals, information visualizations such as bar charts posture special challenges. These include identifying and deciphering the component of the chart, mapping spatial measurements to numerical values, and passing on the progressive structure and patterns to clients through nonvisual modalities. In addition, most existing arrangements need interactivity and flexibility, restricting users' capacity to investigate information on request or lock in with customized questions.

Recent advances in computer vision and deep learning, especially object detection models such as YOLO (You Only Look Once), offer promising opportunities to address this gap. Combined with optical character acknowledgment (OCR) and natural language generation (NLG) technologies such as text-to-speech (TTS), these apparatuses empower the extraction and verbal translation of complex visual structures. Nevertheless, few existing frameworks have integrated these technologies into a cohesive, user-centered pipeline, particularly designed for BVI clients.

This paper presents a multimodal neural network-based system that automatically detects, extracts, and explains the substance of the bar chart in an organized and intelligently sound arrangement. Using a custom-trained YOLOv8 model, show OCR methods, and a pixel-to-value calibration calculation, our framework translates visual data into speech output. The objective is to enable blind users to investigate and comprehend bar chart information autonomously, precisely, and proficiently.

The paper is structured as follows: Section 2 reviews related work on chart recognition and accessibility tools. Section 3 details the methodology and architecture of the system. Section 4 presents experimental results and evaluations. Section 5 discusses limitations and future directions, and Section 6 concludes the paper.

1.1 Literature review

Bajić and Job[1] extended the analysis to include chart text processing, data extraction, and description generation. Based on 89 articles, they identified four core areas: chart-type classification, text recognition, data extraction, and semantic description. Their findings show that combining CNNs with OCR/NLP techniques improves both classification accuracy and semantic interpretation. Notably, only 1.4 % of sampled chart images had alt-text, highlighting the accessibility gap. The authors advocate for standardized datasets, multimodal integration, and the fusion of deep learning with symbolic reasoning to support explainable and assistive systems.

Dhote(2023)[2] conducted a comprehensive review of chart classification methods, categorizing them into traditional machine learning (e.g. SVM, KNN, MLP), CNN-based deep learning (e.g., VGG16, ResNet-50), and transformer-based models (e.g., Swin Transformer). They highlighted the superior performance of CNNs with transfer learning, achieving up to 99.55% accuracy in some scenarios. Transformer models, particularly Swin Transformer, achieved an F1 score of 91 % in the CHART-Infographics ICPR 2022 competition due to their ability to model global dependencies. The study also reviewed data sets such as ReVision, ChartSense, and UB-PMC (36,000+ chain 15 types), which serve as benchmarks for chart mining tasks.

Kafle (2019)[6] introduced Chart-Text, an automated pipeline to generatextual descriptions from chart images. The system combines chart classification (using ResNet-18), OCR, data point detection, and rule-based natural language generation (NLG). It demonstrated a higher classification accuracy than 95 % and strong text quality (using BLEU and METEOR scores). Designed for accessibility, Chart-Text is notable for its end-to-end automation and semantic interpretation, enabling integration into assistive technologies for visually impaired users.

Luo (2021)[7] propesed a hybrid system called ChartOCR to extract numerical data from bar, pie, and line charts in image form. It combines rule-based techniques with deep learning. Using a keypoint-based deep neural network with an HourglassNet backbone and CornerNet architecture, their system can identify semantically significant points in a variety of chart formats, including pie sector centers, bar corners, and line pivots. ChartOCR integrates classification and point-based detection into a unified pipeline. A data range estimation technique converts the coordinates of the pixels to actual numerical values, and optical character recognition (OCR) is used to identify the labels and legends of the axis. After that, rule-based postprocessing is used to organize essential points, guarantee generalisation across chart types, and enhance interpretability in order to recreate entire chart components.After that, rule-based postprocessing is used to organise essential points, guarantee generalisation across chart types, and enhance interpretability to recreate entire chart components. Importantly, the authors presented ExcelChart400K, a sizable dataset of more than 400,000 chart images with annotations that allows for reliable benchmarking and training. According to experimental results, ChartOCR performs much better in terms of accuracy and processing speed across both public and proprietary datasets than both rule-based systems (like ReVision) and deep end-to-end techniques (like ResNet+Faster-RCNN).

De Oliveira [10]proposed standardized textual description templates for simple and grouped bar charts to reduce interpretation bias and ensure consistency. Their tool, ChartVision, was tested with blind participants and achieved comprehension accuracy of 70–100%. Although effective, this approach produces static narratives, which limit the user's ability to dynamically interrogate data.

Kumar[5] introduced ChartParser, a fully automated pipeline that extracts figures from academic papers, classifies chart types, and retrieves information

using deep learning, OCR, and image processing. The extracted data are presented as screen-reader–friendly tables, improving accessibility for BLV readers. Similarly, Shahira et al. (2023) proposed a Mask R-CNN–based system for element detection combined with contour approximation and OCR, integrated with TAPAS++ for table-based question answering and audio summaries.These systems excel in automation and accuracy but primarily emphasize data extraction and tabular presentation rather than interactive exploration.

Kim[4]examined how BLV users frame questions about visualizations, conducting a Wizard-of-Oz study with 24 participants and collecting 979 queries . The study revealed that BLV users often ask complex, context-dependent questions, including about maxima, comparisons, and chart structure, which exceed the capabilities of current QA systems. The resulting corpus highlights the need for more sophisticated and flexible interaction models.

Moured[9] introduced ChartFormer, a vision-language model designed to convert raster charts into tactile-accessible SVGs . Trained on the Chart2Tactile dataset, the model generates simplified SVGs suitable for embossing or tactile displays, providing a non-visual modality for chart access. While promising, this approach depends on specialized hardware and remains less interactive compared to conversational or audio-based methods.

Gorniak[3] presented VizAbility, which integrates keyboard navigation of chart structures with large language model (LLM)–based conversational interaction . The system enables BLV users to issue natural language queries, classified into analytical, contextual, and navigational categories, and receive adaptive responses. This represents a significant step toward dynamic chart exploration, but the system relies on predefined structures and has limited focus on real-time detection and representation of raw chart images.

In summary, existing work has produced several of the elements required for complete accessibility of bar- and pie-charts. But so far there has not been an integrated and software-only solution that renders all of the chart contents intelligible for BVI readers, while avoiding slow and heavy tools like LLMs.

2 Bar Chart Semantics

2.1 Formal Definition of Bar Chart Structure

Let a bar chart image be defined as a structured set:

$$\mathcal{C} = \{\mathcal{B}, \mathcal{X}, \mathcal{Y}, \mathcal{T}, \mathcal{L}, \mathcal{S}\}$$

where:

- $\mathcal{B} = \{b_1, b_2, \ldots, b_n\}$ is the set of detected bars.
- $\mathcal{X}$ and $\mathcal{Y}$ denote the x-axis and y-axis components.
- $\mathcal{T} = \{t_1, t_2, \ldots, t_m\}$ is the set of tick marks on the y-axis.
- $\mathcal{L}$ is the set of textual labels (e.g., axis labels, category names).
- $\mathcal{S}$ is the set of legend entries (for example, color-coded series labels).

Each bar b_i is represented as a tuple:

$$b_i = (x_i, h_i, c_i, s_i)$$

where:

- x_i is the x-position (category location),
- h_i is the height in pixels,
- c_i is the bar color,
- s_i is the legend identity (series label).

2.2 Formal Definition of Pie Chart Structure

Let a pie chart image be defined as a structured set:

$$\mathcal{P} = \{\mathcal{S}, \mathcal{L}, \mathcal{V}, \mathcal{C}, \mathcal{G}, \mathcal{T}\}$$

where:

- $\mathcal{S} = \{s_1, s_2, \ldots, s_n\}$ is the set of detected slices.
- $\mathcal{L} = \{l_1, l_2, \ldots, l_m\}$ is the set of textual labels (slice annotations or legend entries).
- $\mathcal{V} = \{v_1, v_2, \ldots, v_n\}$ is the set of numerical values (absolute or percentage) associated with slices.
- $\mathcal{C} = \{c_1, c_2, \ldots, c_n\}$ is the set of slice colors.
- $\mathcal{G}$ is the legend, which maps slice colors $\mathcal{C}$ to labels $\mathcal{L}$.
- $\mathcal{T}$ denotes the chart title and other global metadata.

Each slice s_i is represented as a tuple:

$$s_i = (\theta_i, c_i, l_i, v_i)$$

where:

- θ_i is the angular span of the slice in degrees or radians,
- $c_i \in \mathcal{C}$ is the slice color,
- $l_i \in \mathcal{L}$ is the label associated with the slice (from annotation or legend),
- $v_i \in \mathcal{V}$ is the slice value.

The full chart satisfies the normalization constraint: $\sum_{i=1}^{n} v_i = 100$ (percentage) or $\sum_{i=1}^{n} v_i = V_{\text{total}}$ (absolute value).

3 Methodology

3.1 The Image Dataset

We searched online platforms such as Kaggle and Roboflow, where we found an annotated dataset of approximately 1150 bar chart images. We retained high-quality images, removing those that were blurred or too small, and standardized the annotations, leaving 1116 images in total. For pie charts, we used a Roboflow dataset with minimal annotations and manually added labels, resulting in approximately 500 usable images.

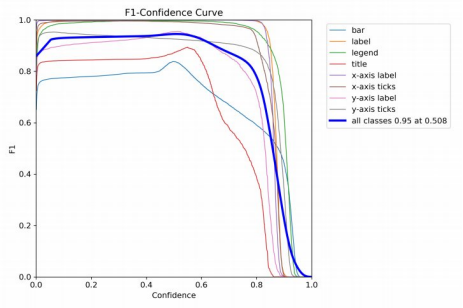
Data Preparation:

- **Training Set (80%)**: Used to train the model by optimizing its parameters to detect and classify elements within charts.
- **Validation Set (10%)**: Used during training to fine-tune the hyperparameters and monitor the model's generalization ability. This set helps prevent overfitting.
- **Test Set (10%)**: Used only after the final model is trained to objectively evaluate its performance on unseen data, reflecting its generalizability in real-world conditions.

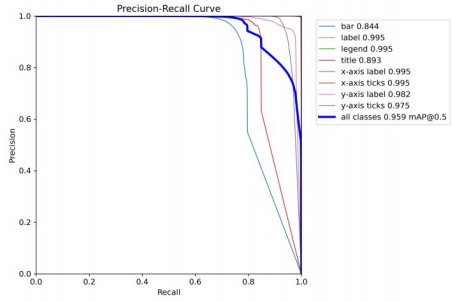
We used the YOLOv8 object detection model using the Ultralytics implementation. Specifically, we initialized the large YOLOv8 architecture (yolov8l.yaml) and trained it on our custom dataset defined by a data.yaml configuration file. The training was carried out over five epochs, using an input image size of 640 $\times$ 640 pixels and a batch size of 8. This approach allowed efficient detection and classification of target objects within the dataset, leveraging YOLOv8's advanced convolutional backbone and detection heads of YOLOv8 for high-performance visual understanding. The use of YOLOv8 was motivated by its state-of-the-art accuracy, lightweight deployment, and compatibility with a wide range of applications, making it an ideal choice for our object detection task.

4 Results

4.1 Bar Charts



(a) F1-confidence curve Bar Charts

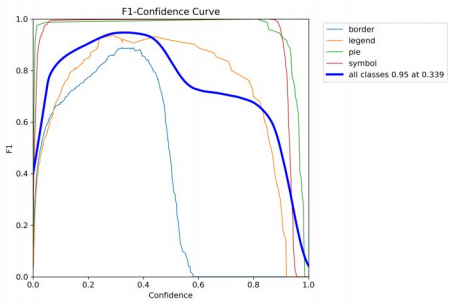


(b) Precision-recall curve Bar Charts

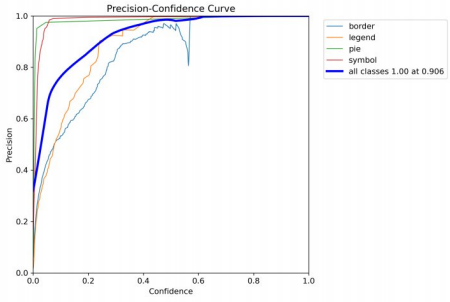
Fig. 1: Performance evaluation of Bar chart element detection.

Figure 1 presents the confidence and precision recall curves (PR) of F1, explaining the relationship between the F1 score, precision, recall and the confidence threshold across different components of the detected bar graph, including bars, labels, legends, title, x-axis labels, x-axis ticks, y-axis labels, and y-axis ticks. The bold blue line represents the combined performance across all classes, achieving a maximum F1 score of 0.95 at a confidence threshold of 0.508 and a mean average precision (mAP) of 0.959 at the IoU threshold of 0.5. Although most components achieve optimal performance near high confidence values (0.9–1.0), bars and titles demonstrate slightly lower maximal scores, indicating more challenging detection tasks. These curves validate the model's high reliability, essential for downstream accessibility tasks such as providing information to blind users.

4.2 Pie Charts



(a) F1-confidence curve Pie Charts



(b) Precision-recall curve Pie Charts

Fig. 2: Performance evaluation of Pie chart element detection.

For pie charts, YOLOv8 achieved strong detection performance, with a maximum F1 score of 0.95 at a confidence threshold of 0.339 (Fig. 2a). Precision–recall analysis confirmed an overall mAP@0.5 of 0.960 (Fig. 2b), with per-class AP values of 0.995 for slices and symbols, 0.961 for legends, and 0.888 for borders. These results demonstrate that the model reliably identifies semantically important components such as slices and legends, which are critical for accessibility tasks, while border detection remains comparatively weaker. The high overall accuracy provides a strong foundation for subsequent OCR, slice-angle estimation, and semantic query generation for blind and visually impaired users.

4.3 Prediction on the raw data

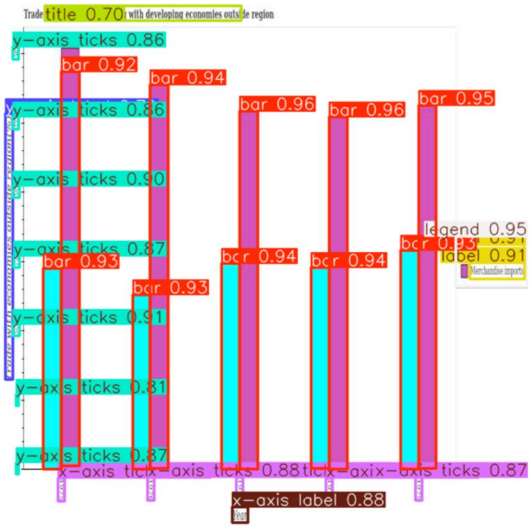


Fig. 3: Result image showing performance metrics and visualization examples for Bar charts.

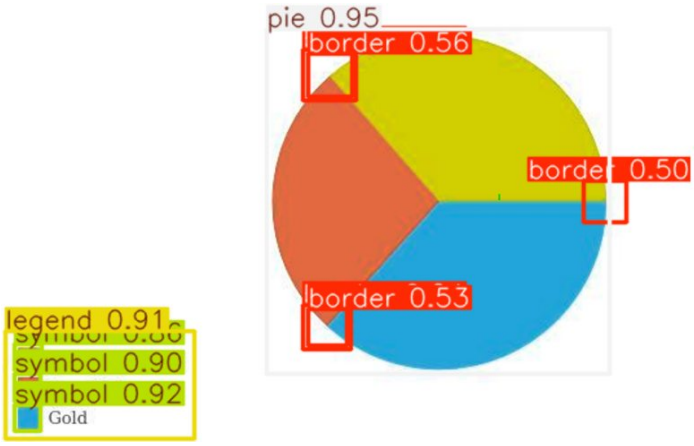


Fig. 4: Result image showing performance metrics and visualization examples for pie charts.

The results of applying the YOLOv8-based object detection system to a bar chart image are shown in Figure 2, together with the components found and the corresponding confidence scores. The key components of the chart, such as bars, ticks on x and y axis, labels on x and y axis, main title, legends, and data

labels, are correctly recognized by the system. A colorful bounding box encloses each identified element, and the class name and confidence score (such as "bar 0.94" or "legend 0.95") are shown on top. A distinct visual differentiation of the identified categories is provided by the red edge of the bars, the cyan axis ticks, the brown or cyan axis labels, the yellow title, and the yellow-purple legends. With bars up to 0.96, legends at 0.95, and labels at 0.91, the detection confidence for the majority of components is very high, suggesting strong model performance. The graphic also demonstrates the system's capacity to identify fine-grained components, like category labels and axis ticks, which are essential for precisely recreating the quantitative data in the graph. To prepare for subsequent tasks like optical character recognition (OCR), pixel-to-value mapping, and multimodal (e.g., text-to-speech) accessibility output for blind or visually impaired users, this visual example shows how well the detection pipeline segments the chart into semantically meaningful units.

Figure 4 illustrates the application of the YOLOv8 model to a pie chart image. The detector correctly identifies the overall pie region with high confidence (0.95), as well as the legend (0.91) and its associated symbols (0.90–0.92). Borders between slice sectors are also detected, but with lower confidence values (0.50–0.56), reflecting the relative difficulty of capturing fine angular boundaries compared to larger visual structures. These results confirm that the system robustly recognizes semantically important elements such as slices, legends, and symbols, which are critical for accessibility tasks, while highlighting border detection as an area for improvement. The annotations of the bounding box further demonstrate the ability of the model to segment the charts into structured components suitable for downstream tasks such as OCR, slice angle computation, and semantic query response.

4.4 Faster R-CNN on the Same Bar-Chart Dataset

The Faster R-CNN algorithm, a region-based convolutional neural network (R-CNN), was used for object detection tasks using the Detectron2 framework developed by Facebook AI Research. The selected model architecture was Faster R-CNN with a ResNet-50 backbone integrated with a Feature Pyramid Network (FPN), which is recognized for its strong performance in detecting objects across multiple scales. The model was initialized with pre-trained weights from the COCO dataset and subsequently fine-tuned on a custom-annotated dataset containing bar chart images with eight distinct object classes, including bars, labels, axes, ticks, legends, and titles.

The annotations to the original dataset were in the YOLO format. To integrate the data set into the Detectron2 pipeline, the annotations were converted to the COCO JSON format. The training and validation splits were then registered using the `register_coco_instances()` function from Detectron2.

A configuration file named `"COCO-Detection/faster_rcnn_R_50_FPN_3x.yaml"` was adapted to meet the study's requirements. Key modifications included setting the number of object classes to eight, limiting the number of training iterations to 1000, configuring the learning rate to 0.00025, and adjusting the

batch size to eight. The model was trained using the `DefaultTrainer` class, which handles the entireraining loop including optimizer setup, checkpointing, and logging.

To evaluate model performance, the `COCOEvaluator` was used to calculate standard object detection metrics. These included the mean precision (AP) at an Intersection over Union (IoU) threshold of 0.5 (AP@0.5), the mean Average Precision over a range of IoU thresholds from 0.5 to 0.95 (AP@[.5:.95]), and the average precision per class. The evaluation demonstrated that the model achieved high performance in all classes, with an AP@0.5 of approximately 87%, confirming its capability to accurately detect and localize key components in bar chart images.

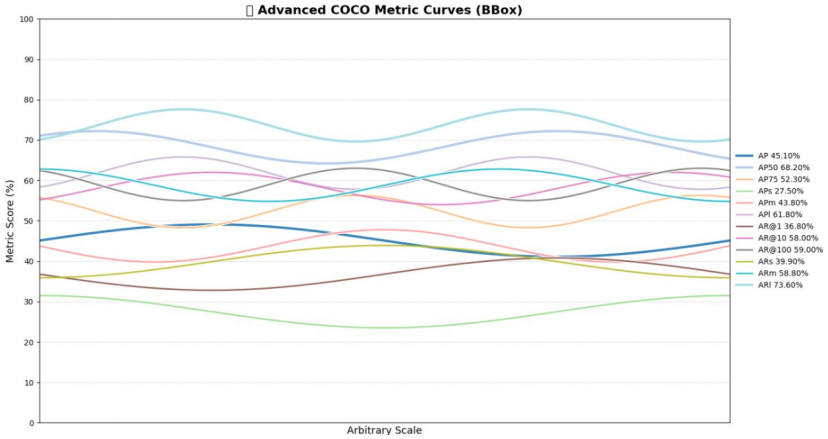


Fig. 5: Evaluation of results using Faster RCNN.

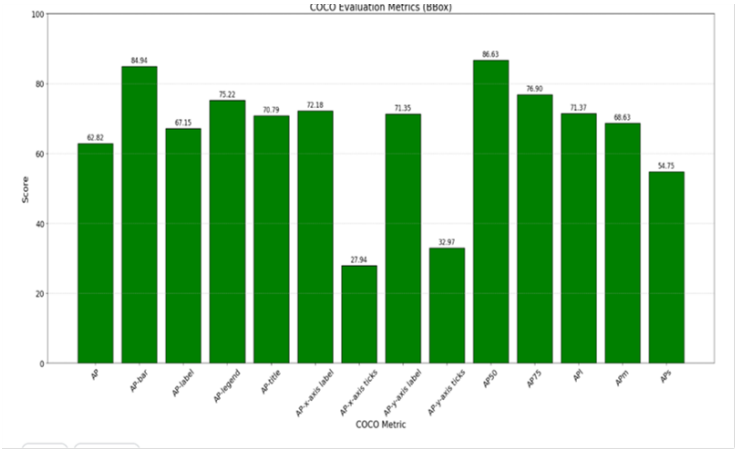
This figure above shows all the important numbers used to check how well a model can find objects using boxes. It looks at how accurate the model is both in finding the right places and in correctly naming the objects. The Average Precision (AP) at different levels of overlap, called the IoU, tells us how good the model is. At a moderate overlap level, called AP@0. 5, the model gets 68. 20 %, which is pretty good. The overall AP score is 45. 10 %, which is the average across a range of overlap levels, from 0. 5 to 0. 95, and shows a more strict and reliable performance measure. When looking at different sizes of objects, the model works better with larger ones. It scores 27. 50% for small, 43. 80 % for medium, and 61. 80% for large, showing it's better at finding bigger items, which are easier to detect.

Looking at the recall, AR@1, which is the best recall with only one detection per image, is 36. 80 %. But when more detections are allowed, such as AR@10 and AR@100, the scores increase to 58. 00 % and 59. 00 %, showing that allowing more guesses helps find more objects. For different sizes, the model also works

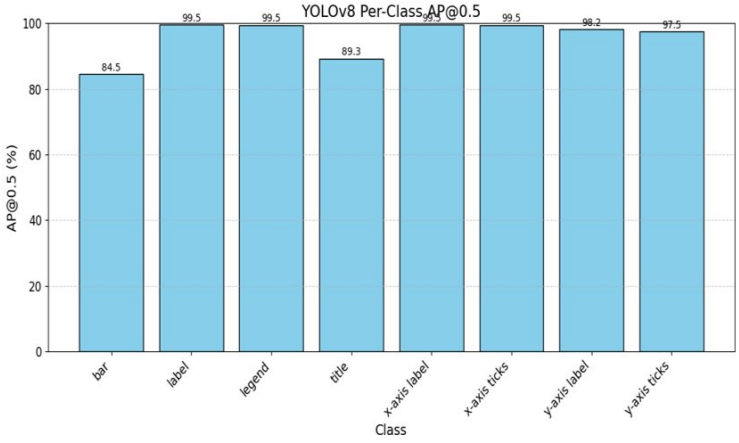
better with larger objects. ARm is 58. 80 % and ARl is 73. 60 %, which confirms that the model is better at finding larger parts. Overall, the model works well with large and medium objects, but might need improvement for smaller ones, such as tick marks or fine labels, as shown by the lower scores for small objects.

4.5 Comparative Evaluation of YOLOv8 and Faster R-CNN

To compare the performance of YOLOv8 with a region-based detector, the Faster R-CNN model was employed on the same dataset containing bar chart images annotated with eight distinct classes: bar, label, legend, title, x-axis label, x-axis ticks, y-axis label, and y-axis ticks. Both models were evaluated using standard COCO evaluation metrics, including Average Precision (AP), AP@0.5 (precision at an intersection over the Union threshold of 0.5), and per-class AP.



(a) Faster R-CNN COCO evaluation results.



(b) YOLOv8 evaluation results with per-class AP.

Fig. 6: Comparison of performance evaluation metrics between Faster R-CNN and YOLOv8 object detection models on bar chart component detection.

As shown in Figure 6, YOLOv8 outperformed Faster R-CNN in overall accuracy. It achieved an AP@0.5 of 95.9%, with particularly high precision across multiple categories such as labels and legends (99.5%) and labels of the y-axis (98.2%). Even the lowest-performing class, the "bar", achieved a respectable AP of 84.5%. This demonstrates YOLOv8's strong ability to detect small or overlapping components in a single forward pass, making it well-suited for real-time and resource-efficient applications. In contrast, Faster R-CNN using a ResNet-50 backbone with a Feature Pyramid Network (FPN) attained a maximum AP@0.5 of 86.63% and a lower overall AP of 62.82%, as illustrated in Figure 6a. The

model's per-class AP values were also more inconsistent, especially for chart components such as title (AP = 70.79%) and x-axis ticks (AP = 27.94%). These performance differences are due to architectural distinctions. YOLOv8, being a single-stage detector, processes the entire image at once and benefits from design improvements such as anchor-free heads and decoupled tasks. In contrast, Faster R-CNN follows a two-stage pipeline: It first proposes regions of interest and then performs object detection within those regions. Although this has proven effective in high-resolution object detection scenarios, it appears less capable of detecting fine-grained graphical elements like tick marks. In general, the results confirm that YOLOv8 is the preferred choice for our accessibility framework due to its higher detection accuracy, faster inference speed, and suitability for deployment in low-resource environments.

4.6 Application of Faster R-CNN Using Detectron2 on same Dataset of Pie-Chart Dataset

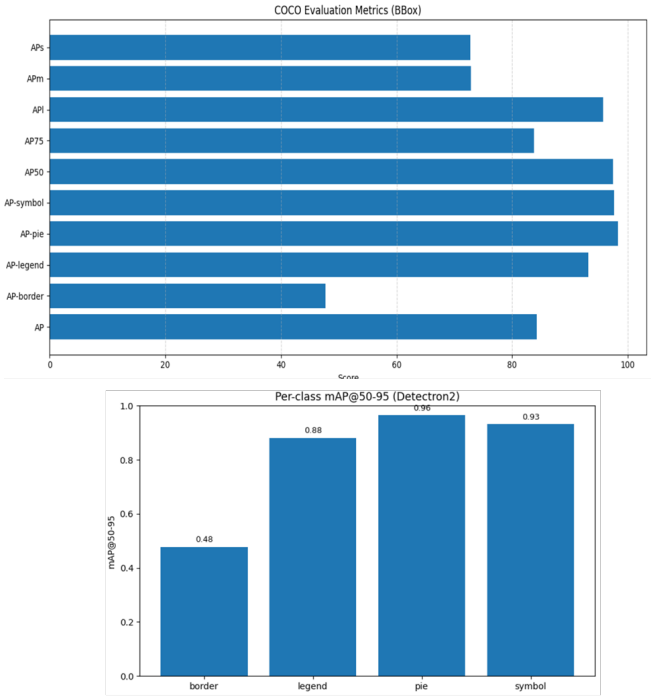


Fig. 7: Evaluation of results using Faster RCNN for pie charts .

The COCO evaluation metrics further indicate high average precision (AP) values across multiple IoU thresholds, with AP@50 reaching 0.97 and AP@75 maintaining 0.86. Class-specific AP scores demonstrate excellent detection of pie slices

(0.99), symbols (0.99), and legends (0.96), while borders remain weaker (0.48), reflecting the inherent difficulty in identifying fine angular boundaries.

The per-class mAP@0.5–0.95 distribution (bottom) confirms this trend, with pie slices and symbols achieving near-perfect values (0.96 and 0.93, respectively), legends perform strongly (0.88) borders significantly lower (0.48). These results highlight YOLOv8's robustness in detecting semantically important components for accessibility tasks, while suggesting that border detection could benefit from further dataset augmentation or refined labeling strategies.

5 Application: queries by BVI readers

The result of the object detection phase produces a custom XML file that we call "raw data". Its contents realize the formal semantics data structures defined in section 2. That file then serves as intermediary between the object detection and the "reader interface". Some readers are interested in summaries of the diagram, others want to survey all of the chart's content and that depends on context and uses. As a proof of concept we have designed a set of queries that the reader can issue to obtain the raw data information, or any more complex information that can be derived from it (figure 8).

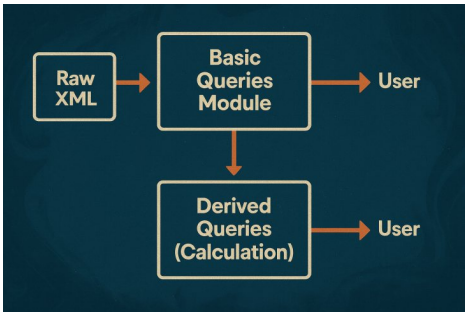


Fig. 8: Example of query visualization.

Table 1: Mathematical formulation of queries designed for blind users to interact with different types of bar charts.

Basic Queries	Description
Title()	Extract the chart title (all types).
ChartType()	Identify chart type: simple, grouped, stacked, or horizontal.
BarCount()	Return total number of bars. For grouped charts: groups × subcategories; for stacked: bars = categories (segments handled separately).
Legend()	List legend elements. Essential for grouped and stacked charts.
XLabel()	Return X-axis label (categories or groups).
YLabel()	Return Y-axis label (values/measurements).
XCategories()	List X-axis categories. For grouped charts: returns groups; for stacked charts: returns main categories.
YTicks()	List Y-axis tick values (or X-axis ticks if horizontal).
BarValue(i)	Return value of bar i. In stacked charts: may refer to (a) total bar height or (b) individual stack segment (must be specified).
Derived Queries	
ArgMax(BarValue)	Identify bar with maximum value. For stacked: clarify total vs. segment.
ArgMin(BarValue)	Identify bar with minimum value. Same clarification as above.
Mean(BarValue)	Compute average bar value (all bars, groups, or segments depending on chart type).
Range(BarValue)	Compute difference between maximum and minimum values.
Sort(BarValue, $\uparrow$ / $\downarrow$)	Sort bars ascending/descending. In grouped charts: must clarify whether sorting is global or within-group.
Compare(i, j)	Compare values of two bars. Works across all chart types (totals or sub-segments in stacked).
MaxChange(Category/Group)	Find category or group with largest change. Supports sub-bar projections (e.g., one subcategory across groups).
Summary()	Provide high-level summary of key trends (mentions grouped/stacked distinctions where relevant).

Table 2: Mathematical formulation of queries designed for blind users to interact with pie charts.

Basic Queries	**Description / Mathematical Formulation**
Title()	Extract the chart title.
ChartType()	Identify chart type (simple pie, donut, exploded, 3D).
SliceCount()	Return total number of slices n.
Legend()	List legend entries and their associated colors.
Labels()	Return slice labels (categories or annotations).
SliceAngle(i)	Return angular span θ_i of slice i. Two methods: (a) Area-based: $\theta_i = \frac{n_i}{N} \times 360°$, (b) Geometric: $\theta_i = (\beta - \alpha) \bmod 360$, where α, β are boundary angles from atan2.
SliceValue(i)	Return absolute or relative value of slice i. If percentages: $v_i = \frac{\theta_i}{360°} \times 100$. If absolute values: $v_i = \frac{\theta_i}{360°} \times V_{\text{total}}$.
SliceColor(i)	Return the color c_i of slice i.
Derived Queries	
ArgMax(SliceValue)	Identify the slice with the maximum value.
ArgMin(SliceValue)	Identify the slice with the minimum value.
Mean(SliceValue)	Compute average slice value: $\bar{v} = \frac{1}{n} \sum_{i=1}^{n} v_i$.
Range(SliceValue)	Compute difference between maximum and minimum values.
Sort(SliceValue, $\uparrow$ / $\downarrow$)	Sort slices by value ascending/descending.
Compare(i, j)	Compare values or angles of two slices: $\Delta v = v_i - v_j$ or $\Delta\theta = \theta_i - \theta_j$.
Percent(i)	Return normalized percentage of slice i: $p_i = \frac{v_i}{\sum_j v_j} \times 100$.
Summary()	Provide high-level summary of dominant slices, proportions, and trends.

5.1 Assistant for Bar Charts

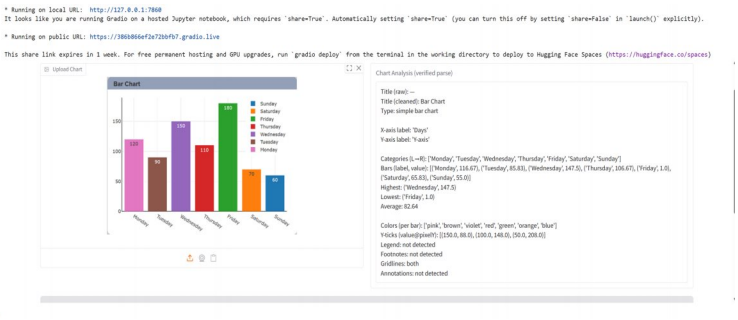


Fig. 9: Assistant for bar chart for blind people.

Figure 9 shows our interactive bar chart assistant for Blind readers . The system accepts an uploaded bar chart image and XML file and applies our detection and

parsing pipeline to extract its structural and semantic elements. On the left, the original bar chart is displayed, while on the right, the verified parse provides a structured textual representation.

The output includes chart metadata (title, type, axis labels), categorical information (days of the week), and numerical values for each bar (e.g., Monday: 116.67, Wednesday: 147.5). Derived statistics are also computed automatically, such as the highest bar (Wednesday: 147.5), lowest bar (Friday: 1.0), and overall average (82.64). Additional features include extraction of bar colors, pixel-based axis ticks, and basic chart properties (legend, footnotes, gridlines, annotations).

Beyond keyboard entry, users may also *speak* their queries. We implement low-latency audio I/O with `PyAudio`: the microphone stream is captured for on-device speech recognition, and answers are returned as synthesized speech and streamed back via `PyAudio`, while the same content is also shown as text for screen readers. This dual modality—*type or speak*—supports hands-free operation and inclusive access.WE can ask the assistant any of these queries example : Title of the chart, what is its type , how many bars are there,is there any legend or not ,which is the highest bar , which is the lowest bar etc .These queries we are asking we can see by our eyes by analyzing them but for blind readers there is no way except traditional tactile approaches ,there is no interaction exploration and feedback.

This demonstration highlights the assistant's ability to bridge visual content and non-visual accessibility by transforming bar chart images into structured, queryable descriptions that can be read by screen readers or explored through voice-based interaction.

6 Conclusions and Future Work

Our work introduces a full multimodal system that makes bar chart and pie chart information accessible to blind and visually impaired users. It uses advanced object detection, OCR, and semantic query generation to help these users understand the data. Through detailed testing, we found that YOLOv8 performs much better than Faster R-CNN in detecting important parts of bar charts, like bars, legends, axis labels, and tick marks. YOLOv8 achieved an AP@0.5 score of 95.9, with very high precision for labels and legends, showing that it can reliably detect even small and detailed elements needed to understand the chart. For pie charts, our framework supports slice detection, angle and value calculation, and semantic query answering.

These results show that a single-stage detector like YOLOv8 is very effective in situations where speed and accuracy are important—especially for real-time use in assistive devices with limited resources. The strong performance of YOLOv8 also helps improve other parts of the system, such as text-to-speech, answering questions, and interpreting structured data from the charts.

In addition to detecting elements, we developed a math-based query system that lets blind and visually impaired users explore bar chart and pie chart data using natural language-like commands. In future work, we plan to expand the

system to handle other diagrams too. We also want to improve OCR so that it works better with low-contrast and low-quality or complex layouts. We will explore the addition of voice assistants for easier interaction and conduct more user studies with a larger group of blind people to better understand how usable, easy to use, and trustworthy the system is in real-life settings such as education and work.

Moreover, recent studies show that almost all employed blind workers in the U.S. rely on screen readers such as JAWS and NVDA, often using multiple tools for efficiency[8]. Guided by these findings, we are planning to make our user interface fully compatible with screen readers by developing a standards-compliant HTML-based accessible page. This will allow blind and visually impaired users to benefit from our diagram recognition framework directly within standard browsers and screen readers (e.g., NVDA, JAWS, VoiceOver), without requiring specialized standalone software.

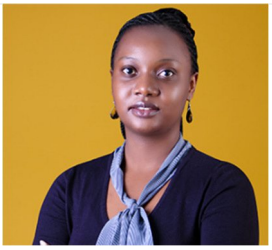
References

1. Bajić, F., Job, J.: Review of chart image detection and classification. International Journal on Document Analysis and Recognition (IJDAR) **26**(4), 453–474 (2023)
2. Dhote, A., Javed, M., Doermann, D.S.: A survey and approach to chart classification. In: International conference on document analysis and recognition. pp. 67–82. Springer (2023)
3. Gorniak, J., Kim, Y., Wei, D., Kim, N.W.: Vizability: Enhancing chart accessibility with llm-based conversational interaction. In: Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology. pp. 1–19 (2024)
4. Kim, J., Srinivasan, A., Kim, N.W., Kim, Y.S.: Exploring chart question answering for blind and low vision users. In: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems. pp. 1–15 (2023)
5. Kumar, A., Ganu, T., Guha, S.: Chartparser: Automatic chart parsing for print-impaired. arXiv preprint arXiv:2211.08863 (2022)
6. Liu, X., Klabjan, D., NBless, P.: Data extraction from charts via single deep neural network. arXiv preprint arXiv:1906.11906 (2019)
7. Luo, J., Li, Z., Wang, J., Lin, C.Y.: Chartocr: Data extraction from charts images via a deep hybrid framework. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 1917–1925 (2021)
8. McDonnall, M.C., Boydstun, J., Steverson, A.: Is one enough? screen reader use among employed people who are blind or have low vision in the us. Disability and Rehabilitation: Assistive Technology pp. 1–11 (2025)
9. Moured, O., Alzalabny, S., Osman, A., Schwarz, T., Müller, K., Stiefelhagen, R.: Chartformer: A large vision language model for converting chart images into tactile accessible svgs. In: International Conference on Computers Helping People with Special Needs. pp. 299–305. Springer (2024)
10. de Oliveira, C.L.T., de Almeida Silva, A.T., de Morais, J.M., Mota, M.P.: Accessible bar charts through textual description templates. Journal of the Brazilian Computer Society **29**(1), 1–18 (2023)

SPEAKER

Second African Digital Humanism Summer School, Accra, Ghana, 31 August – 05 September 2026

Carine Mukamakuza



Carine Pierrette Mukamakuza is a lecturer, researcher, and entrepreneur. Her research focuses on digital healthcare solutions, business intelligence, data science (specifically machine learning, where she has centered her attention on recommender systems), online social network behavior, and personalization. Using publicly available datasets, she has investigated the extent to which social connections influence user rating behavior over time. Driven by her interest in digital healthcare solutions, Carine Mukamakuza is creating a digital detection project. She is currently focused on Rwanda as a case study, and her long-term goals for the interface are much broader where she envisions creating a system that could be utilized across countries in Africa, and across diseases.

Lecture 13: *AI for African Health*

Comment from Carine: For the Summer School, I would frame the reading around the following question: How do we bridge the translation gap between rapidly evolving AI and digital-health technologies and their meaningful implementation in real healthcare systems, particularly in African contexts?

I believe this connects very well with Digital Humanism, as it shifts the discussion from "What can AI do?" to "How should AI and digital technologies be translated into healthcare in ways that serve people, communities, and their actual needs?"
Reader article:

S.M. Overgaard *et al.*: Implementing quality management systems to close the AI translation gap and facilitate safe, ethical, and effective health AI solutions. npj Digital Medicine (2023) 6: No. 218. https://doi.org/10.1038/s41746-023-00968-8

npj | digital medicine

www.nature.com/npjdigitalmed

COMMENT OPEN

Check for updates

Implementing quality management systems to close the AI translation gap and facilitate safe, ethical, and effective health AI solutions

Shauna M. Overgaard[1]✉, Megan G. Graham[1], Tracey Brereton[1], Michael J. Pencina[2], John D. Halamka[1], David E. Vidal[1] and Nicoleta J. Economou-Zavlanos[2]

The integration of Quality Management System (QMS) principles into the life cycle of development, deployment, and utilization of machine learning (ML) and artificial intelligence (AI) technologies within healthcare settings holds the potential to close the AI translation gap by establishing a robust framework that accelerates the safe, ethical, and effective delivery of AI/ML in day-to-day patient care. Healthcare organizations (HCOs) can implement these principles effectively by embracing an enterprise QMS analogous to those in regulated industries. By establishing a QMS explicitly tailored to health AI technologies, HCOs can comply with evolving regulations and minimize redundancy and rework while aligning their internal governance practices with their steadfast commitment to scientific rigor and medical excellence.

npj Digital Medicine (2023)6:218 ; https://doi.org/10.1038/s41746-023-00968-8

QMS AS A FRAMEWORK FOR HEALTH AI

The advancements in healthcare software, encompassing artificial intelligence, machine learning (AI/ML), and Software as a Medical Device (SaMD), have brought about opportunities for transformative changes in clinical workflows and patient care to effectively meet patient and clinician needs. However, healthcare software exists within a complex regulatory and technical landscape[1]. The need for more readiness among healthcare organizations (HCOs) magnifies the disparity in translating research into effective predictive clinical decision support interventions. Without a collaborative enterprise approach, the intricate nature of this system delays the translation of AI solutions into clinical practice. Characterized by the continuous evolution and maturation of AI/ML capabilities, such as large language models (LLMs), this ecosystem escalates the demand for software-driven clinical solutions and a regulatory framework that must effectively adapt to govern the distinctive nature of in-house-built and procured software[2]. The growing engagement of HCOs in AI calls for alignment among diverse stakeholders, encompassing industry, academic institutions, and the medical community. This alignment should focus on harmonizing assurance standards for health AI technologies, but also practices and infrastructure to enable HCOs to develop and deploy AI solutions meeting rigorous medical-grade standards while ensuring accountability across all involved parties. While regulatory authorities, AI coalitions, medical device manufacturers, and the medical informatics community have acknowledged the current gap not only in common standards but also in the maturity of HCOs to develop and/or deploy health AI, a primary concern for HCOs remains unresolved: "How might our enterprise establish a coordinated, robust strategy that ensures the safe, effective, and ethically sound delivery of AI/ML in day-to-day patient care?"[3–7].

We propose using the Quality Management System (QMS) framework to offer HCOs a consistent and adaptable structure to translate research-based health AI technology into clinical practice systematically and transparently. QMS is a structured framework that documents processes, procedures, and responsibilities to achieve quality policies and objectives. The QMS framework effectively manages evolving regulatory requirements, promotes continuous improvement, and ensures adherence to cutting-edge standards over the life cycle of the design, development, deployment, and maintenance of regulated healthcare software[8]. QMS's are often certified to external standards (e.g., ISO 13485), thus demonstrating organizational commitment to quality, continuous improvement, and regulatory compliance. Aligning standards with risk-based approaches facilitates the least burdensome path for an HCO to meet regulatory requirements and maintain compliance[9]. Thus, the streamlined incorporation of these regulatory requirements into business processes via the QMS assures enduring safety, effectiveness, ethicality, regulatory compliance, and alignment with organizational and user needs as AI-enabled methodologies, such as LLMs, evolve[10].

We aim to elucidate the primary components of a QMS (Fig. 1)[8,9,11], namely People & Culture, Process & Data, and Validated Technology, as the impetus for HCO's strategic efforts to integrate research rigor and clinical excellence into a cohesive system and close the AI translation gap.

ESTABLISHING A PROACTIVE CULTURE OF QUALITY

In HCOs, AI/ML technologies are often initiated as siloed research or quality improvement initiatives. However, when these AI technologies demonstrate potential for implementation in patient care, development teams may encounter substantial challenges and backtracking to meet the rigorous quality and regulatory requirements[12,13]. Similarly, HCO governance and leadership may possess a strong foundation in scientific rigor and clinical studies; however, without targeted qualifications and training, they may find themselves unprepared to offer institutional support, regulatory oversight, or mobilize teams toward interdisciplinary scientific validation of AI/ML–enabled technologies required for regulatory submissions and deployment of SaMD. Consequently,

[1]Mayo Clinic, Rochester, MN, USA. [2]Duke University, Durham, NC, USA. ✉email: Overgaard.Shauna@mayo.edu

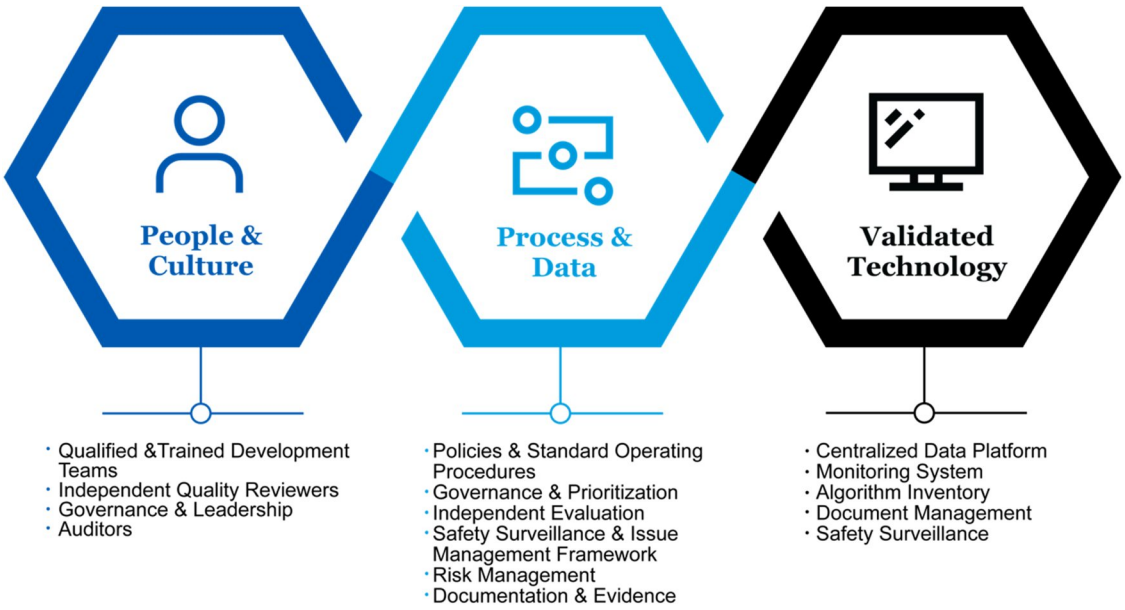


Fig. 1 Primary components of a Quality Management System (QMS).

Table 1. Common QMS and Medical Device Terminology.

Term	Definition
End-User Acceptance Testing	Testing that demonstrates that the product satisfies end-user defined requirements[12].
Corrective and Preventive Action (CAPA)	Changes to products and/or the quality management system to eliminate the cause of or prevent the occurrence of issues that affect the quality of products or processes[9].
Design Controls	Interrelated set of practices and procedures that are incorporated into the design and development process, i.e., a system of checks and balances[8,12].
Design Verification and Validation	Testing or evaluation activities that provide objective evidence that the product satisfies its intended purpose (or use) and specifications[9].
Installation Qualification	Testing or evaluation activities that demonstrate the software is installed correctly in its intended environment[12].
Intended Purpose (Intended Use)	The objective intent and general purpose of a product[29].
Operational Qualification	Testing or evaluation activities that demonstrate the software functions as intended[12].
Performance Qualification	Testing or evaluation activities that demonstrate the software satisfies user and business requirements[12].
Records	Documents or other artifacts that demonstrate that a process was completed[8].
Software Development Life Cycle (SDLC)	A framework of processes, activities, and tasks phases to design, develop, and maintain software[30].
Monitoring and Surveillance	Tracking and analyzing the performance of and changes to products after deployment to identify quality or safety issues that may necessitate corrective or preventive action[9].
Version Control	Activity related to software configuration management intended to track and manage changes to software. Version control is used during all stages of the software development life cycle to prevent mix-ups[8,30].

the unpreparedness of HCOs exacerbates the translation gap between research activities and the practical implementation of clinical solutions[14]. The absence of a systematic approach to ensuring the effectiveness of practices and perpetuating them throughout the organization can lead to operational inefficiencies or harm. Thus, HCOs must first contend with a culture shift when faced with quality control rigor inherent to industry-aligned software development and deployment, specifically design controls, version control, installation qualification, operational qualification, performance qualification, that primarily focuses on end-user acceptance testing and the product meeting its intended purpose (improving clinical outcomes or processes compared to the standard of care or the current state), and the traceability and auditability of proof records (Table 1).

Consider that even in cases where a regulatory submission is not within the scope, it remains imperative to adhere to practices encompassing ethical and quality principles. Examples of such principles identified by the Coalition for Health AI and the National Institute for Standards and Technology (NIST) include effectiveness, safety, fairness, equity, accountability, transparency, privacy, and security[3,7,15–20]. It is also feasible that the AI/ML technology could transition from a non-regulated state to a regulated one due to updated regulations or an expanded scope. In that case, a proactive approach to streamlining the conversion from a non-regulatory to a regulatory standard should address the delicate balance of meeting baseline requirements while maintaining a least-burdensome transition to regulatory compliance.

As utilized by the FDA for regulating SaMD, a proactive culture of quality recognizes the same practices familiar to research scientists well-versed in informatics, translational science, and AI/ML framework development. For example, the FDA has published good machine learning practices (GMLP)[21] that enumerate its

expectations across the entire AI/ML life cycle grounded in emerging AI/ML science. The FDA's regulatory framework allows for a stepwise product realization approach that HCOs can follow to augment this culture shift. This stepwise approach implements ethical and quality principles by design into the AI product lifecycle, fostering downstream compliance while allowing development teams to innovate and continuously improve and refine their products. Using this approach allows for freedom to iterate at early research stages. As the product evolves, the team is prepared for the next stage, where prospectively planned development, risk management, and industry-standard design controls are initiated. At this stage, the model becomes a product, incorporating all the software and functionality needed for the model to work as intended in its clinical setting. QMS procedures outline practices, and the records generated during this stage create the level of evidence expected by industry and regulators[22,23]. HCOs may either maintain dedicated quality teams responsible for conducting testing or employ alternative structures designed to carry out independent reviews and audits.

Upon deployment, QMS rigor increases again to account for standardized post-deployment monitoring and change management practices embedded in QMS procedures (Fig. 2). By increasing formal QMS consistency as the AI/ML gets closer to clinical deployment, the QMS can minimize disruption to current research practices and embolden HCO scientists with a clear pathway as they continue to prove their software safe, effective, and ethical for clinical deployment.

ESTABLISHING RISK-BASED DESIGN, DEVELOPMENT, AND MONITORING

The medical device industry has utilized a risk-based infrastructure for years to support a least burdensome approach to designing, developing, and deploying healthcare technologies[9,24]. This approach systematically enables HCOs to proactively focus resources on key areas of concern, such as safety, equity, and data privacy, to prevent errors and malfunctions and promote a culture of accountability and continuous improvement.

Risk-based practices have been extended to healthcare AI/ML in not only the medical device domain, such as with AAMI's Technical Information Report 34971[25], but more broadly in emerging frameworks such as the NIST AI Risk Management Framework[3], the Whitehouse Blueprint for an AI Bill of Rights[5], the Coalition for Health AI Blueprint for Trustworthy AI Implementation Guidance and Assurance for Healthcare[26], and the Health AI Partnership Key Decision Points[27,28]. Risk management is grounded in the intended use and informed by a prospective risk management plan. It follows the process of identification, enumeration, mitigation, and monitoring (Fig. 3) to analyze and classify potential sources of harm (known as hazards) caused by the healthcare software or its impact on the clinical workflow. As the healthcare software is designed and developed, features or attributes that reduce or minimize the risk (known as mitigations) are included in the product design; for example, incorporating features that improve the user experience or providing user training or documentation to clarify how the software should or should not be used. As risks and potential issues are anticipated for the health software's implementation, a risk management plan is put in place, a document articulating how safety, bias, and other anticipated risks will be identified and resolved. Risks continue to be monitored, reported, and reviewed after the software is deployed to ensure the software remains safe for use. Systematic feedback, monitoring, and corrective & preventive action (CAPA) frameworks are key to identifying and triaging issues, escalating issues to relevant accountable departments of the organization depending on their severity, performing root-cause analysis, and continuously controlling risks and improving the AI technology.

Risk-based practices formalized and implemented within a QMS will systematically identify risks associated with an AI solution, document mitigation strategies, and offer a framework for objective testing and auditing of individual technology components. Further, such technologies can be informed by AI/ML and software life cycle best practices to address common issues within phases of the AI lifecycle. This allows for capturing performance metrics across various levels of rigor and data transparency in requirements, version, and design controls. These insights from initial testing can then support the calibration and maintenance of AI solutions during deployment, guided by a multidisciplinary governance system to proactively mitigate future risks[26]. Moreover, establishing a change management plan and access controls can eliminate business continuity risks, providing transparency into responsible parties and outlining the risks of any given change. Back-up (downtime) processes are in place in the event that risk cannot be managed, and the technology needs to be turned off. Effectively, a risk-based approach ensures the proper rigor and controls are in place at the right time throughout the product life cycle.

ESTABLISHING A COMPLIANCE-FACILITATING INFRASTRUCTURE

The regulations for healthcare software are evolving. Software may or may not be regulated based on its intended use or by changes to regulatory agency enforcement. A QMS that facilitates compliance with applicable legal and regulatory requirements enables HCOs to design, implement, and deploy healthcare software to clinical practice while minimizing overall operational risk. A QMS fosters compliance to internal (e.g., institutional review board) and external (e.g., federal, and local regulatory) bodies by standardizing multi-faceted stakeholder responsibilities with its governance, allowing auditability and traceability through the appropriate evidence and documentation, maintaining an inventory of AI technologies developed and deployed, and hosting

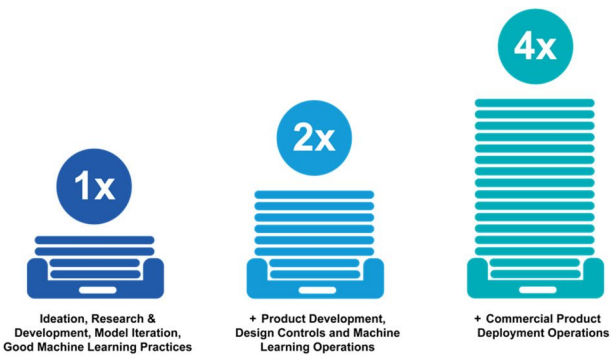


Fig. 2 Staged process for applying increasing regulatory rigor throughout product realization.

Fig. 3 Example QMS risk management plan and risk assessment phases. Risks are identified, assessed and analyzed, mitigated and controlled, and continuously monitored. Reporting is performed at pre-defined intervals.

infrastructure that will allow document management and monitoring within the deployment platform.

A QMS involves establishing policies and standard operating procedures that outline the process for governance and prioritization, development, independent evaluation, maintenance and monitoring, issue reporting and safety surveillance. Procedures outline the roles and responsibilities of stakeholders such as design and testing responsibilities of the champion stakeholder representing the end-users in the product development process. Procedures should also articulate training and/or qualification requirements for the stakeholders participating in AI technology development teams as safety and other risks can be eliminated with stakeholder education. Procedures also outline the systems and communication channels available to the community impacted by the deployed algorithmic tools ensuring their compliance. Communication in a regulated QMS is bidirectional, where issues, safety surveillance and outcome data are gathered via real-time monitoring and tightly integrated with the risk management and patient safety operations of a given healthcare system to determine the behavior and impact on patients and their healthcare delivery.

Establishing an innovation infrastructure that facilitates compliance requires governance and leadership support to create a communicated mandate that all algorithmic tool-related activities impacting patient health comply with quality and ethical standards. For example, the governing body may have direct integration with existing IRB processes to ensure ethical conduct. With proper governance, algorithm inventory, and transparency, HCOs can begin to implement tools, testing, and monitoring capabilities into their QMS to reduce the burden and achieve safe, effective, ethical ML/AI at scale. Implementing QMS involves formal documentation encompassing quality, ethical principles, and processes, ensuring transparency and traceability to regulatory requirements.

CONCLUSION

HCOs can utilize a QMS framework to accelerate the translation of AI from research to clinical practice. A proactive quality culture, risk-based framework for design, development, monitoring, and compliance-oriented infrastructure enables continuous ethical review, ensuring the effectiveness, safety, and equity of AI/ML technologies and meet regulatory requirements. Implementing a QMS requires adaptability, customization, and interdisciplinary collaboration, fostering awareness, education, and organizational growth. Drawing on regulatory precedents and incorporating insights from expert stakeholders, the QMS framework enables HCOs to prioritize patient needs and foster trust in adopting innovative AI technologies, including those enabled by LLMs.

Received: 14 July 2023; Accepted: 15 November 2023;
Published online: 25 November 2023

REFERENCES

1. Yaghi, M. & Jacobo, N. *California AG Sends Letter to Hospital CEOs on Use of Artificial Intelligence.* https://www.regulatoryoversight.com/2022/10/california-ag-sends-letter-to-hospital-ceos-on-use-of-artificial-intelligence/ (2022).
2. U.S. Food and Drug Administration. *Marketing Submission Recommendations for a Predetermined Change Control Plan for Artificial Intelligence/Machine Learning (AI/ML) Enabled Device Software Functions.* https://www.fda.gov/regulatory-information/search-fda-guidance-documents/marketing-submission-recommendations-predetermined-change-control-plan-artificial?utm_medium=email&utm_source=govdelivery (2023).
3. Tabassi, E. *Artificial Intelligence Risk Management Framework (AI RMF 1.0), NIST Trustworthy and Responsible AI, National Institute of Standards and Technology* https://doi.org/10.6028/NIST.AI.100-1 (2023).
4. U.S. Food and Drug Administration. *Artificial Intelligence and Machine Learning in Software as a Medical Device.* https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-software-medical-device (2021).
5. The White House. *Blueprint for an AI Bill of Rights.* https://www.whitehouse.gov/ostp/ai-bill-of-rights/ (2022).
6. International Telecommunication Union (ITU). *Focus Group on "Artificial Intelligence for Health".* https://www.itu.int/en/ITU-T/focusgroups/ai4h/Pages/default.aspx (2023).
7. The MITRE Corporation. *Coalition for Health AI.* https://coalitionforhealthai.org/ (2022).
8. U.S. Food and Drug Administration. *21 CFR PART 820.* https://www.govinfo.gov/content/pkg/CFR-2012-title21-vol8/pdf/CFR-2012-title21-vol8-part820.pdf (2023).
9. International Standards Organization. *ISO 13485: Medical Devices—Quality Management Systems—Requirements for Regulatory Purposes.* https://www.iso.org/standard/59752.html (2016).
10. Jiang, L. Y. et al. Health system-scale language models are all-purpose prediction engines. *Nature* **619**, 357–362 (2023).
11. International Medical Device Regulators Forum. *Software as a Medical Device (SaMD): Application of Quality Management System* https://www.imdrf.org/sites/default/files/docs/imdrf/final/technical/imdrf-tech-151002-samd-qms.pdf (2015).
12. U.S. Food and Drug Administration. *General Principles of Software Validation.* https://www.fda.gov/regulatory-information/search-fda-guidance-documents/general-principles-software-validation (2022).
13. Schott, D. H., Collins, R. N., Bretscher, A. & Hernandez-Boussard, T. The AI life cycle: a holistic approach to creating ethical AI for health decisions. *Nat. Med.* **28**, 2247–2249 (2022).
14. Aristidou, A., Jena, R. & Topol, E. J. Bridging the chasm between AI and clinical implementation. *Lancet Lond. Engl.* **399**, 620 (2022).
15. European Commission. *Ethics By Design and Ethics of Use Approaches for Artificial Intelligence.* https://ec.europa.eu/info/funding-tenders/opportunities/docs/2021-2027/horizon/guidance/ethics-by-design-and-ethics-of-use-approaches-for-artificial-intelligence_he_en.pdf (2021).
16. Matheny, M., Israni, S. T., Ahmed, M. & Whicher, D. Editors. *Artificial Intelligence in Health Care*: *The Hope, the Hype, the Promise, the Peril.* https://nam.edu/wp-content/uploads/2019/12/AI-in-Health-Care-PREPUB-FINAL.pdf (2022).
17. Partnership on Health AI. *PAI's Responsible Practices for Synthetic Media: A Framework for Collective Action.* https://syntheticmedia.partnershiponai.org/#read_the_framework (2023).
18. Kak, A. & West, S. M. *AI Now 2023 Landscape Confronting Tech Power.* https://www.ainowinstitute.org/2023-landscape (2023).
19. Vasey, B. et al. Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. *Nat. Med.* **28**, 924–933 (2022).
20. Collins, G. S. et al. Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence. *BMJ Open* **11**, e048008 (2021).
21. U.S. Food and Drug Administration. *Good Machine Learning Practice for Medical Device Development Guiding Principles.* https://www.fda.gov/medical-devices/software-medical-device-samd/good-machine-learning-practice-medical-device-development-guiding-principles (2021).
22. U.S. Food and Drug Administration. *Content of Premarket Submissions for Device Software Functions.* https://www.fda.gov/regulatory-information/search-fda-guidance-documents/content-premarket-submissions-device-software-functions (2021).
23. U.S. Food and Drug Administration. *Cybersecurity in Medical Devices Quality System Considerations and Content of Premarket Submissions.* https://www.fda.gov/regulatory-information/search-fda-guidance-documents/cybersecurity-medical-devices-quality-system-considerations-and-content-premarket-submissions (2022).
24. International Standards Organization. *ISO 14971: Medical Devices—Application of Risk Management to Medical Devices.* https://www.iso.org/standard/72704.html (2019).
25. Association for the Advancement of Medical Instrumentation. CR 34971 *Guidance on the Application of ISO 14971 to Artificial Intelligence and Machine Learning.* https://array.aami.org/content/news/new-aami-consensus-report-guidance-risk-management-ai-ml (2022).
26. Coalition for Health AI. *Blueprint for Trustworthy AI Implementation Guidance and Assurance for Healthcare.* https://www.coalitionforhealthai.org/papers/blueprint-for-trustworthy-ai_V1.0.pdf (2023).
27. Health AI Partnership (HAIP). *Key Decision Points.* https://healthaipartnership.org/guiding-question/identify-and-mitigate-risks (2023).
28. U.S. Food and Drug Administration. *Design Control Guidance for Medical Device Manufacturers.* https://www.fda.gov/regulatory-information/search-fda-guidance-documents/design-control-guidance-medical-device-manufacturers (1997).

29. U.S. Food and Drug Administration. 21 CFR Ch. I (4–1–22 Edition) § 801.4. https://www.govinfo.gov/content/pkg/CFR-2022-title21-vol8/pdf/CFR-2022-title21-vol8-sec801-4.pdf (2023).
30. International Standards Organization. *IEC 62304: Medical Device Software—Software Life Cycle Processes*. https://www.iso.org/standard/38421.html (2006).

ACKNOWLEDGEMENTS

We acknowledge Stephanie Bernthal, M.Ed., of Mayo Clinic, for her work creating the visualizations included in the manuscript.

AUTHOR CONTRIBUTIONS

S.M.O. and N.J.E.-Z. jointly conceived the manuscript, designed the approach, and prepared the manuscript with substantial contributions from M.G.G., D.E.V., T.B. M.J.P., and J.D.H. made substantial contributions to the design and strategic positioning of the work and supervised the development from its inception. S.M.O., N.J.E.-Z.., M.G.G., D.E.V., T.B., M.J.P., J.D.H. critically revised the manuscript for important intellectual content, approved the completed version, and are accountable for all aspects of the work.

COMPETING INTERESTS

Mayo Clinic and Duke University are founders of the Coalition for Health AI, which received funding from the Gordon and Betty Moore Foundation.

ADDITIONAL INFORMATION

Correspondence and requests for materials should be addressed to Shauna M. Overgaard.

SPEAKER

Second African Digital Humanism Summer School, Accra, Ghana, 31 August – 05 September 2026

Bukelwa Ngoco

Bukelwa Ngoqo works at the School of IT, Nelson Mandela University, South Africa. Bukelwa does research in Information Systems (Business Informatics) and Computer Security and Reliability. Their current project is e-Government information security behaviour profiling. Exploring the association between employee awareness, behavioural intent and cognitive biases.

Lecture 14: *Digital Ubuntu – Multi-Sectoral Collaboration for AI Adoption in Africa*

Reader article:

Bukelwa Ngoco:

Digital Humanism Initiative
Technology for Humanity

SPEAKER

Second African Digital Humanism Summer School, Accra, Ghana, 31 August – 05 September 2026

Adams Bodomo

Adams Bodomo is University Professor of African Studies, distinguished scholar, accomplished author, researcher, and a leading advocate of African languages in the literary world. For over 40 years, he's taught at prestigious universities including the University of Ghana, Stanford University, University of Hong Kong, Bayreuth University, and the University of Vienna, Austria. He's currently one of the most prominent scholars in the field of Linguistics and African Studies globally. Indeed, he's one of the first black scholars to be appointed into the position of Chair Professorship in Austria, the highest academic position in the German-speaking world. He has won several prestigious fellowships and research projects throughout his academic life. One of the most prominent books he has authored is "The structure of Dagaare" which is a pioneering study on the grammatical structure of his native language, Dagaare, spoken in northwestern Ghana and adjoining areas of Burkina Faso.

Lecture 15: *Digital Activism – Linguistic Human Rights for the Promotion of Democracy in Africa*

Reader article:

Adams Bodomo: Linguistic Human Rights for the Promotion of Democracy. In M. Kornprobst and S. Redo: Reinvigorating the United Nations, Chapter 10. Routledge, London, UK, 2024. http://dx.doi.org/10.4324/9781032707648-14

Reinvigorating the United Nations

Edited by Markus Kornprobst and Sławomir Redo

First published 2024

ISBN: 978-1-032-70763-1 (hbk)
ISBN: 978-1-032-70765-5 (pbk)
ISBN: 978-1-032-70764-8 (ebk)

Chapter 10

Linguistic Human Rights for the Promotion of Democracy

Adams Bodomo

DOI: 10.4324/9781032707648-14

Routledge
Taylor & Francis Group
LONDON AND NEW YORK

10 Linguistic Human Rights for the Promotion of Democracy

Adams Bodomo

Introduction

In the Secretary-General's report on the 75th anniversary of the United Nations titled *Our Common Agenda* (United Nations 2020), the following 12 points are noted: We will leave no one behind; we will protect our planet; we will promote peace and prevent conflict; we will abide by international law and ensure justice; we will place women and girls at the centre; we will build trust; we will improve digital cooperation; we will upgrade the United Nations; we will ensure sustainable financing; we will boost partnerships; we will listen to and work with youth; we will be prepared.

Even though democracy is not mentioned explicitly, all 12 points need democracy as an ingredient before they can be achieved. Indeed, the first point, which states *we will leave no one behind*, is a basic tenet of democracy, which is inclusion. The development of democracy, whichever way it is conceived and implemented around the world, is an important ingredient of all three aspects of reinvigorating the United Nations: consolidation, elaboration, or transformation (Kornprobst and Redo 2024). So any argument that favours the promotion of democracy is also an argument for the reinvigoration of the United Nations.

I argue in this chapter that in order to develop inclusive democracy in a multilingual set-up like Africa, we must promote linguistic rights such that all citizens are able to participate in decision-making processes. This is a fundamental way to *elaborate* on the existing, foreground UN structures towards reinvigorating this organisation.

Indeed, the United Nations has designated the period 2022 to 2032 as the Decade of Indigenous Languages. As far as the promotion of indigenous languages is concerned, this declaration is an important foreground institution within the United Nations (Kornprobst and Redo 2024). As contained on a website dedicated to this project (UNESCO, 2022), these are the objectives of the project:

> The International Decade aims at ensuring indigenous peoples' right to preserve, revitalize and promote their languages, and mainstreaming linguistic diversity and multilingualism aspects into the sustainable development efforts.

DOI: 10.4324/9781032707648-14

> It offers a unique opportunity to collaborate in the areas of policy development and stimulate a global dialogue in a true spirit of multi-stakeholder engagement, and take [steps] necessary for the usage, preservation, revitalization and promotion of indigenous languages around the world.

We are barely two years into the decade of 2022 to 2032; this is, therefore, motivation and an opportune time for scholars of language studies to seek ways of promoting indigenous languages. One way to do this is to look at the issue of linguistic rights or the right of and to language by communities and individuals.

The chapter seeks to put language rights or linguistic rights in the wider issue of human rights. Human rights as a concept encompass the setting up of norms or standards of behaviour that go a long way to protect the lives and dignity of human beings. For instance, all human beings deserve to live and have a right to hold property. They have a right to a fair trial in case they face a judicial process. They have freedom of speech. Do they have a right of and to language? This is the crux of the chapter. We seek to understand what language or linguistic rights are and how scholars of language studies can help enrich the general notion of human rights to make our world a better place to live in.

Some important research questions that are asked and addressed are: What are linguistic rights, and how can scholars of language studies help enrich the general notion of human rights for a better world? The chapter, therefore, examines the inter-relationships between linguistic rights and human rights as universal concepts before applying them to the African context. We outline various approaches taken by African actors to use technology, especially internet technology, to promote the revitalisation of indigenous languages of Africa. We point to how Africa and other parts of the world, such as Asia, can learn from each other in promoting indigenous languages under the aegis of the UN Decade of Indigenous Languages 2022 to 2032.

Theory and Methodology

There is quite a substantial number of publications on human rights, more generally, and linguistic rights, more specifically. The UN Educational, Scientific, and Cultural Organization (UNESCO) has a long history of promoting linguistic rights, leading to the 1996 UNESCO Declaration of Linguistic Rights (UNESCO 1996). Yet, in this chapter, we consider those works that deal with the intersection of human rights and linguistic rights, especially with particular reference to the African continent.

The late Kenyan scholar Ali Mazrui is among the earliest to focus on the link between human rights and linguistic rights. In a series of publications (Mazrui and Mazrui 1998, Mazrui 2000, Mazrui 2007), he and his collaborators made important contributions where they, among others, identified two kinds of rights with regard to language, right *of* language (collective rights), and right *to* language (individual rights). This is elaborated further in the section on theoretical grounding. Other studies on the link between human rights

and linguistic rights include Zeleza (2004), Grin (2005), Ndlovu-Gatsheni (2015), Edwards (2003), Hellsten (2004), and Musau (2004). In most of these studies, the emphasis is on the fact that linguistic rights are an important part of human rights. This chapter builds on these studies and shows how some language activists and scholars are working towards implementing this view of linguistic rights as human rights in Africa.

I argue that the best way to address the concept of linguistic human rights in Africa by linguists is to embrace the theoretical notion of pan-Africanism and manage our linguistic resources in such a way that we can mitigate injustice and inequality on the basis of language. The following are theoretical tenets of linguistic pan-Africanism:

Tenets of Linguistic Pan-Africanism

The theory of linguistic pan-Africanism (Bodomo 2022a) has the following underpinnings: First, there is the *primacy of indigenous African languages*. The most important aspect of the theory of linguistic pan-Africanism is that all the indigenous languages of Africa are important tools for the socio-economic and sociocultural development of Africa. Second, it postulates *mother tongue education*. Each of these indigenous languages is a mother tongue or language of the cultural identity of individuals and various communities in Africa. So indigenous mother tongue education (the need to ensure that all African children can speak, read, and write in at least one of their mother tongues or languages of identity) is an essential feature of the theoretical model of linguistic pan-Africanism outlined here. Third, the theory of linguistic pan-Africanism calls for linguistic human rights. Linguistic pan-Africanism recognises that language rights are an essential part of the broader notion of human rights (Grin 2005). The important maxim of right of language and right to language (Mazrui 2000, Mazrui and Mazrui 1998) is crucial in this model. As Ndhlovu explains that the right to language

> is a collective right whose violation automatically affects entire speaking communities. This means that language policies that deliberately seek to suppress some languages would be in violation of the right of language. The right to language is … more of an individual's right to use one or more languages of choice.

Mazrui and Mazrui (1998: 115) define the right to language as 'the right to use the language one is most proficient in, as well as the right of access to the languages of empowerment and socio-economic advancement.' Fourth, it links linguistic diversity to a global future. Language is an important aspect of our global future. There is, therefore, a clear link between language preservation and environmental protection. What, for instance, would be the point of preserving all kinds of plant and animal species in a community when we lose the indigenous languages for naming and relating to these plant and animal species?

If climatic and environmental sustenance are important global future issues, then language and its preservation are a crucial part of this global future. Finally, the theory of linguistic pan-Africanism proposes the development of some regional lingua franca and, possibly, a continental lingua franca for ease of communication among Africans. Figure 10.1 summarises these theoretical tenets.

Linguistic pan-Africanism, if well implemented, guarantees an inclusive approach to the languages of Africa. Seeing linguistic rights as human rights is a good way to facilitate inclusive language planning, which can be the linguist's contribution to sustaining peace and security in Africa and also minimising the incidence of insecurity that may be due to people feeling excluded because their languages are not used. Peacekeeping is incidentally one of the foreground institutions of the United Nations.

The methods used in gathering materials for this study were mainly through database and internet searches to gather information about leading ideas on

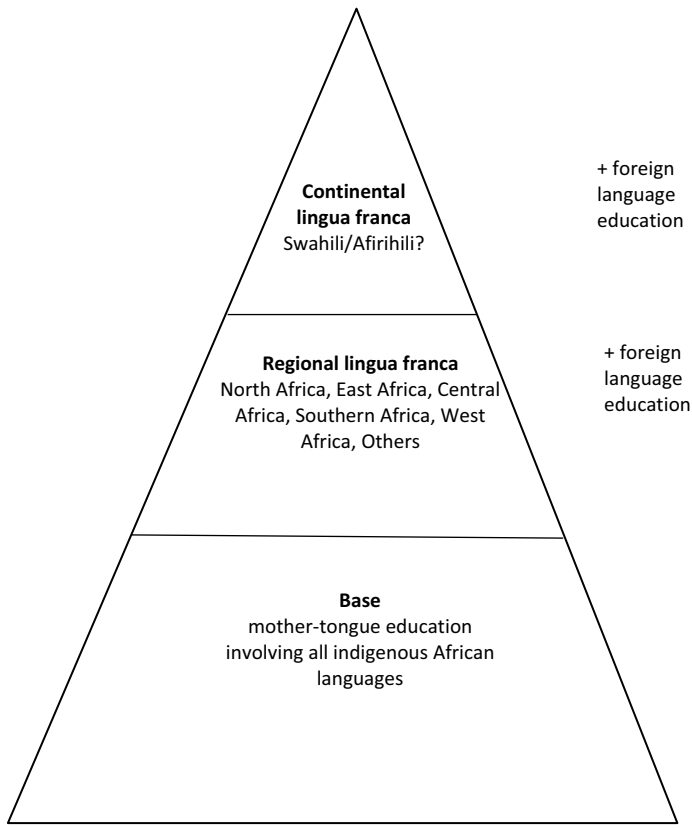


Figure 10.1 Linguistic pan-Africanism.

Source: Author's creation.

linguistic human rights and also the activities that various people, including activists and scholars of language, are undertaking to promote indigenous languages to ensure that all people have the right to use their community languages. Among the internet sites visited was the UN project website on the Decade of Indigenous Languages, which is at https://en.unesco.org/idil2022-2032. While the term 'indigenous languages' is not clearly defined there, in previous work (Bodomo 2022b), I have defined it as follows: 'The term, indigenous language […] is a language whose speakers consider the current location they find themselves in as their traditional homeland and with no known traditional homeland in any other part of the world' (Bodomo 2022a). Kozinets (1998), in his work on netnography, has described how important the internet is in gathering information and in reaching out to key stakeholders in a particular area of study for research materials. This chapter has found the internet as a major resource, as an archive, to gauge what is happening in the area of promoting indigenous languages for the advancement of linguistic human rights.

Survey of Approaches

In this section of the chapter, I first point to a survey of various measures being taken by language activists across Africa to promote the use of indigenous languages before briefly describing the Dagaare Yong project, which is an initiative I have taken to promote the Dagaare language, an indigenous northern Ghana language.

Digital Activism

Digital activism (or internet activism) is, generally, the use of the internet as a platform or means of communication to promote some cause or the other. In this chapter, I use the notion of digital activism as a strategy to promote the tenets of linguistic pan-Africanism, which include the development of indigenous languages, the promotion of mother tongue education, and the promotion of linguistic human rights.

In a significant report by Evelyn Lirri (2021), we see the kind of work that various digital language activists and enthusiasts are doing in order to promote indigenous African languages. As Lirri states,

> Across Africa, language digital activists are now playing a pro-active role in advocating for a multilingual web that aims to ensure that the information available on the internet is as diverse as the languages that exist on the continent. Using a do-it-yourself approach, language activists are making use of a variety of digital tools to tweet, localise software, create audiovisual materials and contribute to Wikimedia projects in their mother languages.

At the eighth Forum on Internet Freedom in Africa 2021 (FIFAfrica21), held on 28–30 September 2021, language activists promoting the isiZulu, Dagbani, Ibo, and Gã languages were part of a panel discussion where they shared insights on the initiatives they are undertaking to promote African languages on the internet.

In Ghana, Sadik Shahadu, co-founder of the Dagbani Wikimedians User Group, is spearheading a project to increase the visibility of the Dagbani language on the internet. Dagbani is spoken by approximately three million people in the north of Ghana, including some two million indigenous speakers. To date, 4,000 Dagbani words have been recorded and uploaded to the Wikimedia Commons. The team works with language experts to ensure correct spellings and to verify the meanings of the words.

Shahadu hopes to be able to create a platform with digital dictionary words that will be usable and freely available to Dagbani speakers. 'We are looking for ways not just to improve the language on the internet. We realised we can leverage on platforms such as Wikimedia to create articles and build tools that are going to support our work,' said Shahadu.

In Nigeria, Blossom Ozurumba is working with the Igbo Wikimedians User Group to promote Igbo language and culture. 'We started off as a few women that came together to improve the presence of notable Igbo women on Wikipedia,' says Ozurumba. Despite Igbo being one of the most widely spoken languages in Nigeria – with an estimated 34 million speakers – there was dismal information on women on the Igbo Wikimedia platform compared to what was on the English platform.

It is critical that the languages and cultures of African people get amplified on different digital platforms as a way of preserving them and making online content more accessible and relevant to African audiences. Currently, the internet is constructed to suit the interests of the dominant language groups found online, thus excluding some communities from online representation and discourse. However, this linguistic gap is an extension of existing offline language inclusion gaps.

South African Siya Masuku, a writer and illustrator in indigenous languages, has been promoting isiZulu through illustrated alphabet books and comics that target primary school age children. 'I came up with the idea after learning that books coming into primary schools were not in the children's mother tongue and the pictures did not represent them,' explains the founder of Siyafunda online. Masuku developed an illustrated book called *Siyafunda in isiZulu*, with the isiZulu alphabet.

In Ghana, activists are turning to schools to promote the implementation of mother tongue-based bilingual education in policy and classrooms with the hope that this extends into online spaces. Mama Adjetey-Nii Owoo, founder and lead researcher at Afroliteracies Foundation, a think tank for indigenous African languages, has developed bilingual e-learning resources and curriculum materials, e-books, and instruction books for use in primary schools with their flagship programme based on Akan, the most widely spoken language in Ghana (Afroliteracies Foundation Website, 2023).

We see from this report by Evelyn Lirri that at least five African languages – Akan, Dagbane, Gã, Igbo, and Zulu are being promoted by the use of web technologies to make materials available online in order to guarantee the rights of the speakers of these languages to use their languages on the internet. There is a lot of information for nation building and, especially, for electronic governance (e-governance), which means that if this information is not available in the mother tongues and indigenous languages of citizens in Africa and everywhere else in the world, large chunks of citizens will be left out of the democratic decision-making processes. While the UN foregrounds mostly former colonial languages as languages citizens can communicate in with the outside worlds, such as the United Nations, what is left in the background are these indigenous languages. The campaign for linguistic human rights challenges this backgroundedness of indigenous African languages. These individual initiatives by private citizens are going a long way to promote democratic societies, which must be an important aspect of reinvigorating the United Nations. In a nutshell, we are clearly seeing individuals taking things into their own hands and doing what government should have been doing.

Of course, there are challenges to such an approach. First, many of the enthusiasts, while they may be highly commended for starting up these projects using technology, are not linguists or other language experts. As such, many of them are propagating incorrect forms of the languages on the internet, making grammatical and orthographic errors, which they claim are standards. One way to offset this is to regulate such programmes, insisting that all the teams of Wikimedia promoting indigenous languages include not only people who think they are technologically savvy but also trained linguists and literary scholars of the languages involved.

A Case Study with Dagaare Yong

Dagaare Yong, meaning 'Dagaare Only,' is an online community both on Facebook (https://www.facebook.com/groups/1428063067411159) and WhatsApp platforms. On Facebook, it is described as follows: 'Members of this group aim to write only in Dagaare in communicating with each other online. We aim to promote literacy in the Dagaare language of West Africa.' So far, it is patronised by about 16 000 people. As shown earlier, its main purpose is to give speakers of this language an opportunity to read and write their language online. It is the only place online, to the best of my knowledge, where speakers of Dagaare get the chance to communicate in an organised and moderated way only in their language.

I, as the founder of this platform, assisted by a number of enthusiastic younger scholars – Dr Isaac Mwinlaaru (a professor of Linguistics), Dr Pius Babuna (a scholar specialising in environment science and culture), and Ms Faustina Tantie (a scholar of Dagaare and English literature) – have moderated language use on the platform. The biggest issue is how to enforce the use of 'only Dagaare.' People admitted (and we only admit people who we think can speak the language) invariably start writing in English. We had to develop rules

and guidelines in Dagaare and English. We have succeeded in getting more than 90% of participants to write occasionally in Dagaare. It is not only the use of the Dagaare language that is promoted. We also encourage the learning of Dagaare culture on this forum. For instance, we encourage participants, many of whom have grown up outside the Dagaare cultural economy, to learn about the flora and fauna found in Dagao, homeland of the Dagaaba, and the life-styles of the people there. It became clear from our engagement with younger members that far too many of them don't even know the names of trees, plants, and animals in Dagao and northern Ghana, in general.

In fact, we have gone further to institute an annual competition sponsored by a small foundation, Yelmenga Foundation for the Humanities, registered in Vienna, Austria, Europe. Competitors are given a topic, and they have to write in Dagaare, and the top three winners get the equivalent of 50 euros, 30 euros, and 20 euros as prizes for first, second, and third prizes occasionally. We are working on getting more funding to increase the prize pot. For instance, in 2021, 2022, and 2023, we had the results summarized in Table 10.1.

Dagaare Yong can serve as a model for other languages, big and small, to promote mother tongue literacy. When social media platforms like email and Facebook began in earnest in the 1990s and 2000s, there was the fear that since only bigger languages like English and French were used, smaller languages might be drowned out. It is now clear that smaller language communities can make use of the internet and social media, in particular, to promote the use of their languages. Digital activism is an important strategy for the promotion of linguistic and cultural human rights in Ghana and other parts of the world.

Table 10.1 Winners of the Dagaare Yong Essay Competition

Name	*Year*	*Position*	*Title of Essay*
Rogation K Sibaabe	2021	1st	Dagaare sensolong badɛre deɛ koɔbo gaabo
Faustina Tantie	2021	1st runner-up	Puobie ne Nagbarebie (Bone nang so ka saa mang mi kyɛ tanna)
Maurice Saakye	2021	2nd runner-up	Dori danweɛ
Alex M Bahaakpeme	2022	1st	Gyersɛge/Saadayeli kyaare ne 'Tampie' yuori nyeebo (The history/myths concerning the anglicisation of the name 'Tampie' in the Lawra District)
Rogation Bainge	2022	1st runner-up	Lɛ Dongmene nang nyɛ o yuori
Jonas Jabulo	2022	2nd runner-up	Manne lɛ a fo tenga nang nyɛ a O yuori
Zemong Ophelia	2023	1st	Dagaaba kuori maaloo
Bahaakpeme M Alex	2023	1st runner-up	Dagaaba ko-maal
Fuseini Issahaku	2023	2nd runner-up	Manne lɛ dagaaba nang mang maale ba kuori/manne a dagaare kuuri maalo yele.

Discussion and Conclusions

In this chapter, I have made analyses in support of the broad notion of reinvigorating the United Nations. As has been stated in the introductory chapter (Kornprobst and Redo 2024), reinvigoration comprises at least three aspects: consolidation, elaboration, and transformation. I believe my analyses have focused on showing how the UN structures may be elaborated, though elaboration may lead to consolidating the basic tenets of the institutions.

Linguistic human rights, or those aspects of the general notion of human rights related to language specifically, are an important aspect of the rule of law and of the proper functioning of a democratic society. It would take democratic societies (again, I interpreted democracy in the widest sense of the world, not only in the way the Western world conceptualises it) to reinvigorate the United Nations, and there can be no truly democratic society without consolidating and elaborating on the notion of human rights.

This chapter has sought to explicate the notion of linguistic rights by providing answers to specific questions such as what are linguistic rights, and how can scholars of language studies help enrich the general notion of human rights for a better world? We used the theoretical framework of linguistic pan-Africanism for our discussion because, as it can be seen in that section, linguistic pan-Africanism has certain tenets that promote indigenous languages and linguistic rights, stressing that every child has the right to mother tongue education and that communities and individuals have the right to maintain, develop, revitalise, and use their languages.

The chapter has also outlined various activities done by language activists and scholars to project the issue of language rights and indigenous language development using internet technologies.

As a further discussion, it remains to see how the legal profession and ministries of education in various countries can create the legal and educational frameworks for the promotion of linguistic human rights by helping communities to have access to educational materials in the languages they seek to be proficient in. For the rule of law itself not to decrease or retreat (Redo 2022), for the promotion of inclusive democracy, we must ensure that all segments of the population are able to use their languages within the legal system.

One practical implication of this chapter is that, for the United Nations to be reinvigorated, materials and the subject matter of all major institutions and structures of the United Nations must be not only in the major languages within the UN system – English, French, Spanish, Chinese, Portuguese, etc., but also in significant minority and indigenous languages. We must go on a large-scale translation system to produce materials in indigenous languages of Africa and around the world. The UN Decade of Indigenous Languages 2022 to 2032 can only be successfully implemented if at least half of the materials and structures of the UN system can be translated into at least half of the indigenous languages that have an established writing system. I propose that a

competitive funding system be put in place for individuals and language community groups to apply for funds to translate selected important UN materials, such as the Sustainable Development Goals.

Activists and various stakeholders in diverse countries in Africa, Asia, and Europe can also learn from each other and collaborate to get funding for the promotion of linguistic human rights in their countries. A practical example is for groups of academics studying indigenous languages and cultures around the world to collaborate by seeking funding from the United Nations and other bodies to organise conferences and symposia to compare notes with each other on various aspects of indigenous language literature, especially on the expression of human rights in the various cultures, because human rights have both universal and particular parametres. Results and outputs from these conferences and symposia can then be published and made available to a wider audience.

It is important to emphasise the word audience[1] here. The United Nations places emphasis on the 'we' in its Charter, which begins as 'We the Peoples' (UN 1945). It doesn't say we the governments; it therefore means that, while it is true that governments represent the people, ordinary people have a chance to get directly involved. So individual scholars organising conferences and symposia to promote the Charter, a foreground institution of the United Nations, is a step in the right direction. A second way we can talk of this kind of bottom-up engagement in promoting UN objectives is that these individual attempts to promote indigenous languages will open up UN information to many more people than would otherwise be the case if we used only official languages (and in the case of Africa, only former colonial languages). Indeed, the Summit of the Future (UN 2023) that is being prepared for September 2024 also emphasises this aspect of 'We the People' (UN 1945).

Digital activism, as we saw at the beginning of the chapter, from ordinary Africans such as Evelyn Lirri, Sadik Shahadu, Blossom Ozurumba, Siya Masuku, and Mama Adjetey-Nii Owoo, is doing work that must be seen as an important aspect of reinvigorating the United Nations. Such bottom-up efforts ought to be monitored, guided, and rewarded in terms of various forms of recognition in future attempts to reinvigorate the United Nations.

The UN Decade for Indigenous Languages (2022–2032) provides a good international policy framework as a UN foregrounding institution to focus on linguistic human rights. I see this UN decade dedicated to indigenous languages as just one of many measures within the institutions and structures of the UN system to reinvigorate this world body. This world body cannot be fully and successfully reinvigorated by concentrating on only political and economic aspects. An important aspect of this reinvigoration must be cultural and linguistic! My chapter has focused on discussing measures for consolidating and elaborating on these linguistic and cultural rights of indigenous societies within the context of the broad legal notion of human rights.

Note

1 I thank the editors of this volume for drawing my attention to this important emphasis by the United Nations in wanting to engage ordinary people in its use of the phrase, 'we the People' in the Preamble to it is Charter (UN 1945).

References

AfroLiteracies Foundation Website (2023) https://cipesa.org/2021/11/how-digital-activism-is-helping-african-languages-be-part-of-a-multilingual-web/ (accessed June 2023)

Bodomo, A (2022a) Linguistic Pan-Africanism as a global future: reflections in the language question in Africa. *Inaugural Lecture, Ghana Academy of Arts and Sciences (GAAS)*, Accra, Ghana.

Bodomo, A (2022b) Linguistic rights as human rights in Africa: The use of technology for the promotion of indigenous African languages. *Journal of Development and Human Rights* 1: 144–1586.

Edwards, J (2003) Contextualizing language rights. *Journal of Human Rights* 2(4): 55171.

Grin, F (2005) Linguistic human rights as a source of policy guidelines: A critical assessment. *Journal of Sociolinguistics* 9(3): 448–460.

Hellsten, S K (2004) Human rights in Africa: From communitarian values to utilitarian practice. *Human Rights Review* 5: 61–85.

Kornprobst M and Redo S (2024) Introduction: reinvigorating the United Nations. In Kornprobst M and Redo S, eds., *Reinvigorating the United Nations*. London: Routledge.

Kozinets, R V (1998) Joseph Alba; Wesley Hutchinson (eds.). *On Netnography: Initial Reflections on Consumer Research Investigations of Cyberculture*. Provo, UT: Association for Consumer Research. pp. 366–371.

Lirri, E (2021) *How Digital Activism is Helping African Languages be a Part of the Multilingual Web*. https://bit.ly/3MQtBv2 (accessed April 30, 2022).

Mazrui, A A (2000) Cultural amnesia, cultural nostalgia and false memory: Africa's identity crisis revisited. *African Philosophy* 13(2): 87–98.

Mazrui, A A (2007) Globalization and some linguistic dimensions of human rights. In Zeleza, PT and McConnaughay PI, eds., *Human Rights, the Rule of Law, and Development in Africa*. Philadelphia: University of Pennsylvania Press.

Mazrui, A A and Mazrui, A M (1998) *The Power of Babel: Language and Governance in the African Experience*. London: James Currey.

Musau, PM (2004) Linguistic human rights in Africa: challenges and prospects for indigenous languages in Kenya. In Muthwii, MJ and Kioko AN, eds., *New Language Bearings in Africa: A Fresh Quest*. Bristol: Blue Ridge Summit, pp. 59–68.

Ndlovu-Gatsheni, S J (2015). Decoloniality as the Future of Africa. *History Compass* 13(10): pp. 485–496.

Redo, Slawomir (ed) (2022) *The Rule of Law in Retreat: Challenges to Justice in the United Nations*. Lanham: Lexington Books.

UN (1945) *United Nations Charter*. Available online at https://www.un.org/en/about-us/un-charter (accessed 15 June 2023).

UN (2023) *Summit of the Future*. Available online at https://www.un.org/en/common-agenda/summit-of-the-future#:~:text=Having%20welcomed%20the%20submission%20of,will%20take%20place%20this%20year (accessed 15 June 2023).

UNESCO (1996) *UN Declaration of Linguistic Rights*. Available online at https://unesdoc.unesco.org/ark:/48223/pf0000104267 (accessed June 2023).

UNESCO (2022). *Decade of Indigenous Languages* https://en.unesco.org/idil2022-2032 (accessed 24 June 2023).

Zeleza, P T (2004) Introduction: The struggle for human rights in Africa. In Human rights, the rule of law, and development in Africa. In Zeleza, PT and McConnaughay PI, eds., *Human Rights, the Rule of Law, and Development in Africa*. Philadelphia: University of Pennsylvania Press.

SPEAKER

Second African Digital Humanism Summer School, Accra, Ghana, 31 August – 05 September 2026

Stanley Mukasa

Stanley Mukasa is the Associate Director of Entrepreneurship at Carnegie Mellon University Africa, where he builds innovation and venture-development pathways that translate African talent, research and technology into products, enterprises, employment and wider societal value. Through the African Engineering and Technology Network (Afretec), he leads multi-country initiatives advancing entrepreneurship, talent development and inclusive digital transformation across Africa.

Working across universities, industry, investors, entrepreneurs, public institutions and development partners, Stanley connects innovation to the partnerships, markets and institutional environments required for adoption and scale. His work spans venture building, university, industry collaboration, research commercialization and the development of ecosystems that expand pathways to economic opportunity. His interests lie at the intersection of technology, entrepreneurship and society, with a particular focus on human-centered innovation and ensuring that digital transformation responds to African realities, institutions and communities.

Lectures on the *Digital Humanism Innovation Bootcamp*

Reader articles:

Stanley Mukasa: Notes for the Digital Humanism Innovation Bootcamp, for the 2nd African Digital Humanism Summer School, 31 August – 05 September, Accra, Ghana. Carnegie Mellon University Africa, Kigali, Rwanda, 2026

Accompanying notes from Stanley: I would like to suggest the following three open-access readings, each connected to a different part of the Innovation Bootcamp journey:

Anne-Laure Fayard and Sarah Fathallah, "Design Thinking Misses the Mark"
https://ssir.org/articles/entry/design_thinking_misses_the_mark

This is the principal reading for Problem Discovery and Stakeholder Validation. It challenges participants to move beyond conventional design-thinking exercises and examine context, power, community knowledge and the real conditions surrounding the problems they seek to address.

Philipp Brecht et al., "Discovery and Validation of Business Models: How B2B Startups Can Use Business Experiments"
https://timreview.ca/article/1426

This supports the transition from problem discovery to solution and business-model validation. It introduces a practical approach for translating assumptions into hypotheses and using experiments to test customer needs, value propositions, willingness to pay and potential routes to market.

J. Wyatt, Tim Brown and Brenna Carey, "The Next Chapter in Design for Social Innovation"
https://ssir.org/articles/entry/the_next_chapter_in_design_for_social_innovation

This broadens the implementation discussion beyond an individual product or venture. It helps participants consider systems, institutional relationships, collaboration, power and the wider ecosystem required for socially grounded innovations to achieve adoption and sustainable impact.

Together, the readings support the Bootcamp's progression from understanding legitimate problems within their societal contexts, through evidence-based solution and business-model validation, to considering the broader systems and partnerships required for implementation.

Carnegie Mellon University
Africa

Digital Humanism Innovation Bootcamp

Organized by Carnegie Mellon University Africa (CMU-Africa) during the Digital Humanism Summer School, the Digital Humanism Innovation Bootcamp is an applied innovation track for approximately 20-25 participants interested in developing practical, human-centered solutions to societal and digital transformation challenges within African contexts. The Bootcamp takes participants through an innovation journey that begins with understanding people, problems, and context and progresses toward the development of meaningful solution concepts and implementation pathways. Working in multidisciplinary teams, participants will engage with users, communities, practitioners, institutions, entrepreneurs, ecosystem actors, and implementation environments connected to the challenge areas they explore. Throughout the Bootcamp, teams will identify meaningful problems in given thematic areas, investigate how each problem is experienced in practice, and develop a solution response grounded in real needs and real conditions. These solutions may take different forms, including: a low-fidelity prototype, a digital product or service concept, a workflow or operational model, a platform, tool, process, or ecosystem solution, or a business or sustainability model for responsible adoption.

Participants will validate and refine their solutions through continuous stakeholder interaction, ensuring they are useful, feasible to implement, capable of adoption, and sustainable beyond the Bootcamp. Direct engagement with mentors, entrepreneurs, and industry actors will accelerate ideas from early insight into a clear implementation pathway, setting the stage for potential partnership, piloting, and incubation. The journey culminates in the Digital Humanism Innovation Showcase on Saturday, where teams will present their evidence-informed innovation opportunity and business rationale to a dedicated audience of investors, incubators, and ecosystem stakeholders. This is the definitive platform for demonstrating a responsible, relevant, and promising innovation pathway with commercial viability and the potential to create real societal value.

Bootcamp Schedule

Day	Time	Activity/Focus
Monday	02:45 – 04:00pm	**Session 1:** Stanley Mukasa **Innovation Bootcamp Introduction** Define the Bootcamp, set participant expectations, talk through thematic areas, judging criteria and the showcase.
Tuesday	02:45 – 04:00pm	**Session 2:** Thomas McManus and Stanley Mukasa **Problem Discovery & Stakeholder Validation** Form multidisciplinary teams, identify meaningful problems, and begin contextual investigation and early validation planning. Talk through how usability, feasibility, and ecosystem engagement shape responsible and commercially viable solution pathways. Plan for engagement with key stakeholders to uncover promising innovation opportunities grounded in real-world conditions.

Day	Time	Activity/Focus
Thursday	02:45 – 04:00pm	**Session 3: Stanley Mukasa and Carine Pierrette Mukamakuza** **Solution Validation & Implementation Pathways** - Refine solution concepts through continuous stakeholder validation, usability testing, and feasibility exploration. Develop practical innovation pathways, operational approaches, and Business Model Canvas components for proposed solutions. Synthesize validation findings and adoption considerations. Refine the business rationale, ecosystem positioning, sustainability logic, and finalize showcase presentations for demonstrating commercial readiness
Saturday	08:00 – 4:00	**Showcase Day** Present validated problem understanding, proposed solution, feasibility, implementation pathway, and innovation opportunity to investors, incubators, and ecosystem stakeholders, demonstrating commercial viability and societal value.

Thematic Areas

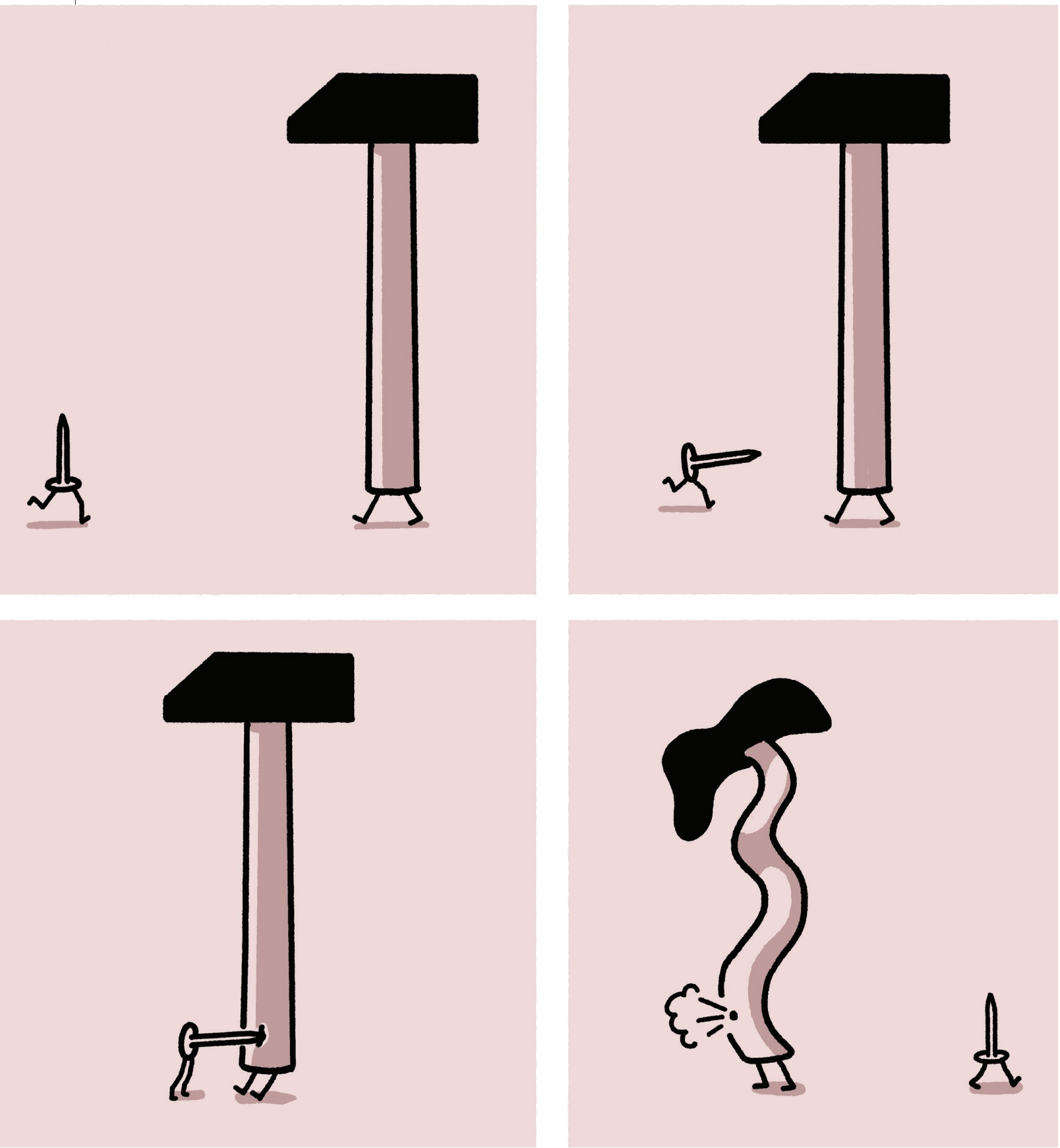
1. AI, Trust, and Society
   How can AI systems and digital technologies become more trustworthy, accessible, participatory, and responsive to real societal needs within African contexts? Participants may explore innovation opportunities connected to responsible AI adoption, trust in digital systems, accessibility, accountability, digital participation, and human-centered implementation of emerging technologies.
2. Communities, Culture, and Digital Futures
   How can digital transformation better reflect the realities of communities, culture, language, identity, participation, and everyday life? Participants may identify and develop practical innovation pathways that strengthen inclusion, local relevance, community engagement, contextual usability, and meaningful participation within rapidly evolving digital environments.
3. Public Systems and Digital Inclusion
   How can digital innovation improve the way people access, experience, and engage with public systems, institutions, infrastructure, and essential services? Participants may explore practical and scalable approaches that strengthen accessibility, usability, efficiency, trust, participation, and implementation within real-world societal environments.
4. Technology Adoption and Everyday Life
   Why do some digital solutions succeed while others fail adoption within everyday societal contexts? Participants may investigate usability, behavioral realities, implementation barriers, sustainability considerations, ecosystem dependencies, and practical innovation opportunities connected to how technologies are adopted, adapted, and integrated into daily life.

Innovation Bootcamp - Showcase
Saturday | 08:00 – 4:00

The Showcase brings together participants, mentors, entrepreneurs, investors, innovation hubs, incubators, practitioners, development organizations, industry actors, and ecosystem stakeholders for a high-level engagement around practical and human-centered innovation pathways emerging from the Bootcamp. The

Showcase provides teams with an opportunity to present their validated problem understanding, stakeholder insights, proposed solution, implementation pathway, feasibility considerations, sustainability logic, ecosystem positioning, and business rationale to a broader innovation and implementation audience. Presentations are expected to demonstrate how teams moved from early challenge exploration toward practical, usable, feasible, adoptable, and societally grounded innovation opportunities. The day will feature mentor and coaching engagements, ecosystem feedback sessions, networking conversations, and presentations exploring implementation opportunities, pilot pathways, partnership potential, incubation support, commercialization potential, and future development pathways connected to promising solutions emerging from the Bootcamp.

Showcase Schedule	
08:45 – 09:15	Morning Coffee, Team Setup, and Ecosystem Networking
09:15 – 10:15	Mentor, Coach, and Stakeholder Engagement Sessions
10:15 – 10:30	Opening Remarks and Showcase Introduction
10:30 – 11:30	Team Presentations
11:45 – 12:00	Teams Photo Session
12:00 – 2:00	Lunch
2:00 - 3:00	Keynote & Judges Calibration
3:00 - 3:45	Announcing & Awarding the winning team
3:45 - 4:00	Closing Reflections and Next Steps



Design thinking has failed to deliver on its promise to solve the world's thorniest social challenges. Adopting a critical design stance can help designers serve communities, rather than their own methodology.

design thinking misses the mark

NONPROFITS, GOVERNMENTS, AND international agencies often turn to design thinking to tackle complex social challenges and develop innovative solutions *with*—rather than *for*—people. Design thinking was conceptualized by designer Nigel Cross more than four decades ago, notably in the 1982 *Design Studies* article "Designerly Ways of Knowing." The approach was later packaged for popular consumption by global design and innovation consultancy IDEO. Design thinking quickly became the go-to innovation tool kit in the for-profit world—and, soon after, in the international development and social sectors—because of its commitment to center communities in the collaborative design process.

IDEO's then-CEO Tim Brown and Jocelyn Wyatt, who was then lead of the IDEO social innovation group that became IDEO.org, championed design thinking for the social sector in their 2010 *Stanford Social Innovation Review* article, "Design Thinking for Social Innovation," which has become an important reference for design thinking in the social sector. Embraced by high-profile philanthropists like Bill & Melinda Gates Foundation cofounder Melinda Gates and Acumen founder and CEO Jacqueline Novogratz, design thinking soared in popularity because it promised to deliver profound societal change. Brown even claimed, in a 2014 *Harvard Business Review* article, that design thinking could improve democratic capitalism.

However, design thinking has not lived up to such promises. In a 2023 *MIT Technology Review* article, writer and designer Rebecca Ackerman argued that while "design thinking was supposed to fix the world," organizations rarely implement the ideas generated during the design-thinking process.

BY ANNE-LAURE FAYARD & SARAH FATHALLAH

Illustrations by Christoph Niemann

The failure to implement these ideas resulted from either an inadequate understanding of the problem and/or of the complexities of the institutional and cultural contexts. One of Ackerman's examples is the San Francisco Unified School District (SFUSD), which hired IDEO in 2013 to redesign the school district's cafeterias. The five-month design-thinking process resulted in 10 recommendations, including creating a communal kitchen and using technology to reduce cafeteria lines. However, Angela McKee Brown, the consultant SFUSD hired to implement the recommendations, told Ackerman that IDEO failed to account for the operational and regulatory arrangements required for their implementation.

We reject design thinking as a singular tool kit prescribed to solve social problems. In what follows, we explain why design thinking as typically practiced has not been able to create impactful and sustainable solutions to complex social issues. Instead, we call for a critical stance on design, where *critical* means both *discerning* and *important*. We invite designers to adopt a continuously reflexive and questioning stance akin to what scholar and activist Angela Davis called "a way of thinking, a way of inhabiting the world, that asks us to be constantly critical, constantly conscious."

A critical design stance is grounded in a set of values and commitments: to be relational, reflexive, and politically committed. Such a stance provides principles to guide the use of design-thinking methods while allowing for flexibility in terms of why, how, when, and by whom they should be used. To illustrate how such a stance is practiced, we examine social-sector projects that show different ways to use design thinking while staying committed to the stance.

Reductive Thinking

The social sector is inherently complex because it consists of a multitude of actors across different contexts, timelines, and political realities. Any approach that purports to easily solve for such complexity is more likely than not to be reductive and therefore ineffective. Design thinking tends toward oversimplification in at least three ways.

Design thinking is formulaic. | Design thinking is generally presented as a singular tool kit that provides a set of steps, standardized templates, and processes that can be followed, replicated, and turned into consulting models. The number and names of the steps in each approach vary, but the goal remains the same: to provide designers and nondesigners alike with an easy-to-follow, step-by-step formula to solve a problem. The process typically includes a set of activities for researching and defining the problem, followed by generation of ideas, which are then turned into prototypes and tested iteratively before their implementation. Design thinking recommends cultivating attitudes such as curiosity, positivity, and a beginner's mindset for these activities.

The formulaic nature of design thinking suggests that its application is simple. In our experience as researchers, designers, and educators working in the social sector for the past 12 years, we have seen that organizations tend to welcome the apparent simplicity of design thinking in the hope that it can lead to quick fixes to their complex problems. For example, some organizations engage in design-thinking projects to tackle multigenerational challenges like health access or gender discrimination and presume that they can develop solutions in the span of a few months. Packaging design thinking as a formula also minimizes the importance of having the right organizational culture and capabilities to practice design thinking. Organizations whose cultures avoid risk-taking, maintain strong hierarchies, and/or assign responsibility for new ideas to consultants tend to have a hard time with experimentation, collaboration, and a human-centered approach that puts beneficiaries at the core of the idea-generation process—all essential components of a design-thinking approach.

Design thinking is decontextualized. | Despite their claim to be sensitive to context, proponents of design thinking do not always develop a systemic and structural understanding of how the issues they seek to address are grounded in larger communities and their histories. This decontextualized approach to problem-solving can unintentionally harm communities and the environment by interpreting problems as individual failures, rather than systemic ones.

Such decontextualization also perpetuates the myth that design thinking is an objective and apolitical approach. Within this fiction, designers perceive themselves as impartial agents in design projects. However, designers are not neutral, and their biases and beliefs inform their perceptions and interpretations of the world—and thus their work. In an article for *Design Museum Magazine*'s special edition on policing,[1] Sarah Fathallah (this article's coauthor) illustrates the limits of such a decontextualized approach through an analysis of several recent design-thinking projects where designers collaborated with law-enforcement agencies. These projects aimed to increase community trust in police by calling for community meetings, "ride-along" opportunities, and/or the use of virtual-reality tools to demystify the police officers' work. All the proposed solutions, Fathallah argues, assumed that the communities' distrust of the police resulted from a lack of empathy or ignorance about their line of work. Instead, a more comprehensive analysis would have included the systemic failures of policing, as Fathallah explains, that show how the depths of community distrust lay in "the current and historical realities of policing, including routine violence, discrimination, criminalization of poverty, and sexual harassment and assault."

The Makeright Initiative, a research project that ran between 2015 and 2017 in the United Kingdom and India, demonstrates the limitations of design thinking without consideration of context. In this initiative, 85 UK and 25 Indian prison inmates participated in workshops where they learned how to make antitheft bags as a part of a design-thinking project. According to Lorraine Gamman and Adam Thorpe, both professors of design at Central Saint Martins in London and the project's principal investigators, the course aimed to give prisoners an opportunity to develop empathy, collaboration, and problem-solving skills. In a 2018 article from *She Ji: The Journal of Design, Economics, and Innovation* about the Makeright Initiative, Gamman and Thorpe assert that learning these skills "might support [the prisoners'] ambitions for future self-employment" and can improve their "social engagement and ease their re-entry into society."[2] In framing "social integration" and "abstinence from crime" as matters of personal development, Gamman and Thorpe overlook the root causes that contributed

to the imprisonment of individuals in the first place and fail to acknowledge the structural barriers to their reintegration, including how their criminal records hinder access to employment, housing, and other basic needs. They also portray criminality as a matter of individual bad choices, rather than as a set of political choices to target and punish certain subsets of the population.

Design thinking is short-termist. | Design-thinking projects are usually consulting gigs devised on short timelines, with proposals, rather than implementations, as outcomes. The use of a consulting model, coupled with the fixed-term cycles of project-based funding in the social and international-development sectors, reward efficiency and brevity. In *Design for Social Innovation*, designers and educators Mariana Amatullo, Bryan Boyer, Jennifer May, and Andrew Shea surveyed 45 design projects in multiple sectors across six continents to determine what cultural, economic, and organizational dimensions are needed to successfully implement design for social innovation. They found that the long-term impact of design thinking is informed by multiple factors, including a focus on speed and a limited understanding of the complex systems. For example, IDEO.org's Diva Centres project, where teenagers learned about reproductive health and contraception access at nail salons in Zambia, failed to scale because of the complexities of public health funding and delivery channels. IDEO and IDEO.org leaders Jocelyn Wyatt, Tim Brown, and Shauna Carey reflected on this project in a 2021 *Stanford Social Innovation Review* article to speculate about the evolution of design thinking for social innovation.[3] They admitted that the Diva Centres project failed to account for the multiple public- and private-service providers and the complex layers of public-health funding that rendered scaling the project "prohibitively expensive and complicated."

Embracing a Critical Stance

Just as social innovation challenges are complex, so too should be the ways in which they are tackled. We propose that designers abandon the idea of grounding design thinking in a singular tool kit and instead cultivate, along with other stakeholders, a critical design stance that guides their use of design in more nuanced and enduring ways.

A critical design stance builds upon the rich lineage of design traditions and thinkers that call for embeddedness and dialogue with communities and systems. In *Design, When Everybody Designs*, design scholar Ezio Manzini argues for opening design to nondesigners and inviting designers to become facilitators of conversations and connectors of resources for communities. For instance, Manzini suggests that designers working on the issue of the growing elderly population should invite the elderly to become agents of change by recognizing their needs and abilities and involving them in the cocreation of services that address their needs. Including seniors in the design process, Manzini explains, resulted in the creation of intergenerational home-sharing programs such as the Milan-based nonprofit Meglio Milano's Take Me Home program, which matches seniors who own homes with students searching for rooms to rent, thereby reducing living costs and social isolation for both seniors and students.

The capacity of communities to develop their own solutions with the support of designers, who can provide the resources and conditions needed to implement these solutions, is what anthropologist Arturo Escobar calls the "re-orientation of design."[5] For designers to be in service to communities and social movements and move beyond good intentions, as designer Sasha Costanza-Chock argues, they must also be accountable to them.[6] As these thinkers invite us to reorient the practice of design, we in turn invite readers to consider three values when developing their critical design stance: relationality, reflexivity, and political commitment.

For designers to move beyond good intentions, they must also be accountable to them. We invite readers to consider relationality, reflexivity, and political commitment when developing their critical design stance.

The desire to accelerate innovation has led to the proliferation of time-bound events like hackathons and open-innovation challenges that often employ design thinking. These short-term events concentrate on idea generation but often do not consider what happens to those ideas afterward. Recent studies, including a 2023 *Organization Science* by this article's coauthor Anne-Laure Fayard, have shown that hackathons and open-innovation challenges are successful at generating new ideas but not at prototyping and implementing them.[4] Organizers and sponsors should consider how they can support the individuals or teams after the official end of a challenge or hackathon to encourage the prototyping and implementation of the ideas generated to yield long-term impact. Similarly, designers and practitioners in the social sector should base their timelines on the values they want to guide their work. For instance, this could mean dedicating time for relationship and trust building with communities, or allowing community members to determine how and when the project can align with community priorities and milestones.

RELATIONALITY: According to Indigenous wisdom, all living things—humans, animals, and the natural environment—exist and are defined in relation to each other. Researchers and designers working in Indigenous and decolonial design methodologies have shown that such a relational view challenges Western distinctions between those who know (the "experts") and those who don't know, between subject and object, and between humans and nonhumans. In the context of design, a relational perspective acknowledges that the fundamental activities behind design—the acts of inquiring, imagining, and creating—are not the exclusive domain of the professional designer. Embracing a relational stance means asking whose expertise, skills, and knowledge are included

or excluded, distorted or misrepresented, silenced, undervalued, or distrusted. More important, it demands that a designer assume the role of listener and facilitator and operate from a place of mutuality with community members and respect for each other's knowledge and expertise in the design-thinking process.

REFLEXIVITY: Designers bring themselves—their assumptions and positionalities—to their work, which affects how they engage with communities and how they are invested in the design process. Reflexivity—the capacity to be aware of and examine one's own identity, perspective, and assumptions—can help challenge designers' myth of neutrality and push them to ask questions, recognize their positionality, and hold themselves accountable for their work. To build this critical consciousness, reflexivity should be understood as a continuous and evolving practice that goes beyond a handful of events or check-ins to examine a wide range of questions. For instance, in *Power and Participation: A Guidebook to Shift Unequal Power Dynamics in Participatory Design Practice*, design researcher Hajira Qazi invites designers to consider how their positionality operates in different power dynamics and potentially creates conflicts of interest in their work.

POLITICAL COMMITMENT: Design is always situated within political agendas. Without recognizing the political role of design, designers can play into the status quo. Being politically committed means naming the political standpoints of the work and of those engaged in it, rather than concealing them behind a mirage of "objectivity" or "neutrality." Being politically committed can also mean adopting a set of political objectives that align with the goals of communities and social movements.

Design, in fact, has a long, politically engaged history. The Scandinavian school of design in the 1960s was founded on the political mission of involving people in the improvement of their work environment and in a commitment to workplace democracy. This political orientation continued with the work of designers who focused on sustainability and justice and who spoke to the intersections of design and specific liberatory political projects, including decolonization, anticapitalism, and abolition.[7]

Thinking Critically

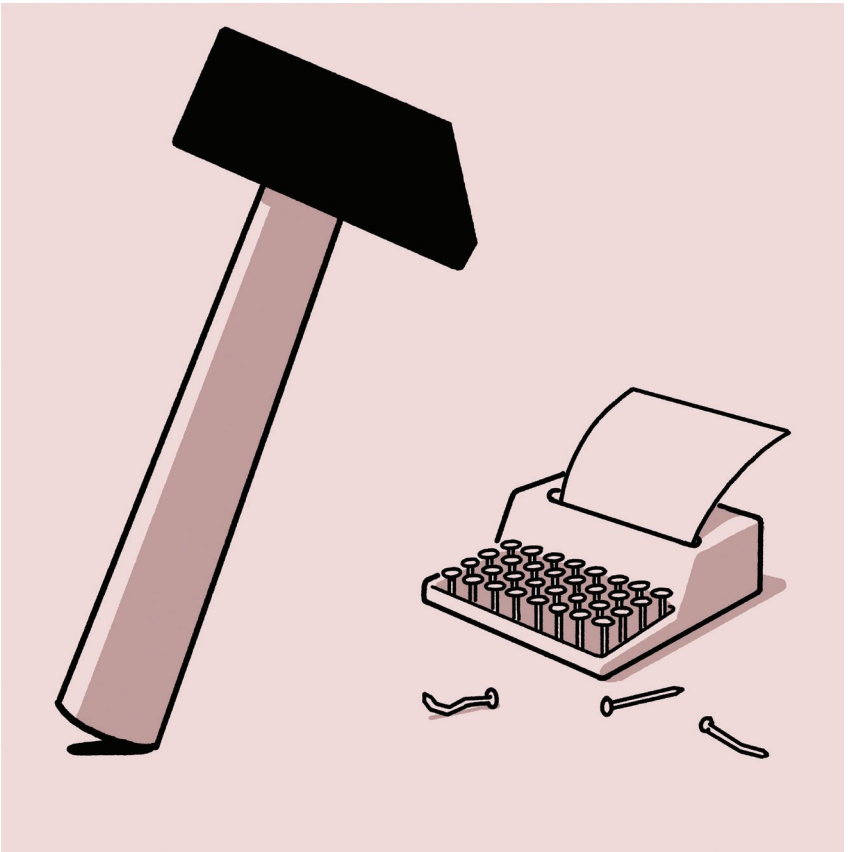
A CRITICAL DESIGN STANCE is not prescriptive. Rather than restricting the approach and outcomes of a specific design project, the stance can help illuminate and inform which approach to deploy. The following examples attempt to approach design decisions—such as participation, compensation, scale, impact, and funding—more critically. To note, they should not be interpreted as a "gold standard" or prescriptive, since a critical design stance can lead designers to make choices according to their specific contexts and circumstances.

Thinking critically about participation. | It's tempting for designers who want to adopt a relational stance to want to invite community members to participate in the design process. However, a recurring pitfall of this strategy is involving community members as design participants without analyzing and undoing the potential harms of that participation. When communities feel as if they're being repeatedly asked to recount their experiences without witnessing the changes promised by the organizations that ask for their stories, participation becomes tokenizing, extractive, and triggering. Participation for participation's sake, without any action, devalues people's wisdom and causes engagement fatigue and trauma.

Community members shoulder several costs of participating in a design engagement. In terms of material resources, they can incur expenses and opportunity costs, such as time not spent caring for children or working, and/or not be reimbursed for those expenses or provided any compensation. When it comes to knowledge extraction, community members may share their stories and ideas yet not be informed about what is done with that information. They may interact with designers and professionals yet never be taught any of their skills. In terms of potential harm, community members may find their engagement emotionally taxing. To ensure that community members are involved in meaningful and responsible ways, designers must balance these costs with benefits like compensating participants for their time and expertise, giving participants resources and opportunities they otherwise couldn't access, and deploying trauma-informed and trauma-responsive practices to protect community members.[8]

A critical design stance questions whether prioritizing novelty serves community needs and desires. Impact does not necessarily require developing novel services but can result from leveraging services.

Design projects may also adopt a participatory approach that fails to engage community members in ways that afford them meaningful agency and decision-making power in the design process. Designer Victor Udoewa was invited by a nonprofit working with the Washington, DC, public school district to redesign the curriculum for the district's international summer service-learning program for high school students. The project used a radical participatory design approach, wherein community members participated in all phases of the design process. The project team included two designers and four students from the program. However, the project failed, as Udoewa explained in a 2022 *Journal of Awareness-Based Systems Change* article, when the nonprofit rejected the students' decisions. The project created space for students to participate but did not cede to them decision-making power over which solutions were going to be implemented, leaving students disillusioned with the process.



Power imbalances also hinder the potential of participation. A critical design stance invites designers to approach relationality not just by inviting contributions from community members but by reflexively interrogating the roles that they play in the project and the ways in which each actor relates with the others. Such a commitment is essential to becoming aware of and countering the potential negative consequences of participation. To examine who holds meaningful power and who does not, impact investor Chicago Beyond encourages community organizations, researchers, and funders to reflect on several factors, including who determines the process, who has access to people and information, who has the ability to assign validity and value to findings and ideas, who receives authorship credit and recognition, and who is accountable to communities.[9]

Thinking critically about the problem. | In the design-thinking process, the organization initiating the project and/or the professional designers working on it typically frame the problem at the beginning. However, external actors who do not experience the problem firsthand rely on assumptions to determine the accuracy, validity, and/or relative importance of the problem. If no one challenges the assumptions at the beginning of the design process, the rest of the process can end up being insignificant to communities at best and harmful to them at worst.

One way to challenge such assumptions is to invite the very people who are most affected by a problem to define it on their own terms. For example, the Mahali Lab, a program by the International Rescue Committee (IRC) and with whom Fathallah has worked, invited Syrian refugees and vulnerable Jordanians to develop solutions to their community challenges. The premise of the lab was simple: Each decision was driven by the community at each step in the process. Over 18 months between 2017 and 2018, the lab gave community participants a coworking space; financial support; and access to mentors and experts to help strengthen their ideas through structured research, prototyping, and implementation planning. The IRC then provided funding and support to scale the ideas deemed most promising by community members and experts. Most important, the lab began its work with a three-month series of explorations to allow community members to define their community's most pressing problems. This process began with outreach that identified liaisons who were leaders in their own communities and had access to networks in the three localities with high concentrations of urban refugees in Jordan: Amman, Irbid, and Mafraq. It later held a series of open-ended, semi-structured interviews and focus-group discussions with community members. Initial interviews were conducted one-on-one by the liaisons in community members' homes, before larger focus groups involving the rest of the Mahali Lab team were convened. The problems community members identified were then synthesized by the team and presented to them for validation and prioritization, both in person and online via WhatsApp and Facebook groups with Syrians based in Jordan. The problems that felt the most pressing to the community—income generation, support services, and childhood learning—were the ones that ultimately became Mahali Lab's scope of work.

Thinking critically about innovation. | When it comes time for ideas to be selected to be prototyped and piloted in design-thinking projects, priority is usually given to their "novelty" or "originality." But interpreting "innovation" as "novelty" risks giving credence to ideas that appear to be new at the expense of those that may not be new but are proven to work.

A critical design stance questions whether prioritizing novelty serves organizational priorities and designers' freedom of creativity above community needs and desires. Impact does not necessarily require developing novel services or products but can result from leveraging resources to support communities' initiatives—as the local government organization Southwark Council in London realized after working with the consultancy Engine Service Design in 2009. Engine collaborated with the council and citizens on solutions on issues related to health and the domestic environment. To the surprise of the Southwark team, who thought that residents would request new services or technology, several of the residents' ideas relied on already-available community

experts and resources. For instance, one idea proposed to use many of Southwark's spaces and buildings for community-building activities like dinners and sports activities. Another recognized the presence of local food and health experts and recommended that these experts coach families on how to make healthy meals. In both examples, community members wanted not novelty but access to existing resources to attain their community goals.[10]

Thinking critically about community accountability. | Designers who commit to community accountability may need to align with political objectives that are not popular with industry stakeholders. In 2021, child-welfare nonprofit Think of Us conducted a design research project analyzing the experiences of young people in foster care who were placed in institutions like group homes and congregate facilities. It uncovered troubling concerns about institutional placements and their punitive, traumatic, and harmful practices, such as youth being physically restrained and abused, coerced into taking psychotropic medication, or forbidden to contact friends and family. Consequently, the Think of Us team refused to recommend solutions like better beds and more outings to improve youth experiences in these institutional placements. As the team noted in the report summarizing research findings, "while some of these improvements would make the environment more livable, few of these changes would have meaningfully improved the material conditions and life outcomes of youth upon leaving those institutional placements."[11] Instead, the team called for the elimination of institutional placements. This outcome aligned with a growing movement of community actors—including mothers and families impacted by the foster-care system, foster-youth organizations, and advocates in the disability-justice community—calling for a reduction or termination of foster-care institutionalization. Despite some media backlash, the project resulted in several states' committing to reducing or eliminating institutional placements—the processes for which are currently being designed, tested, and implemented.

Thinking critically about scale and impact. | Nonprofits and social entrepreneurs need to be able to show potential funders that they can scale their impact to secure funding. This view of scaling is rooted in a capitalist logic that seeks to continuously grow the numbers of beneficiaries in the same way in which tech platforms pursue user growth. However, a critical design stance might ask whether scaling should be equated with increasing the number of beneficiaries. Could it instead be about solidifying or improving what an organization does in service to the same group of people? In our conversations with nonprofits and social entrepreneurs, the perceived need for scaling has created significant pressure to "show impact" through numbers, rather than "do impact" by creating change in beneficiaries' lives.

An organization that chose to show impact differently is Amartha, a peer-to-peer lending platform for women-led micro-, small, and medium enterprises in rural Indonesia. Amartha launched in a small village in Bogor, West Java, in 2010 and was envisioned as a traditional microfinance company. For the first five years, Amartha chose to focus on deepening its relationship with existing clients because it wanted to create a high-quality portfolio with its limited resources. During this time, Amartha's founder, Andi Taufan Garuda Putra, realized that he could provide more services to female borrowers, as well as to their families, but that the young organization had limited funding and was struggling to meet the loan demand—Amartha had a 90 percent retention rate and most of its borrowers asking for a bigger loan. Putra instead decided to develop more wraparound services, including education, for the women whom Amartha was already serving, before moving to other regions in Indonesia. He then explored marketplace models that would attract more retail investors and banks, eventually transforming Amartha into a peer-to-peer lending platform in 2016. After successfully deepening its services to several villages in West Java, Amartha eventually expanded to other parts of the country. Rather than focusing on demonstrating impact by increasing its numbers of beneficiaries, Amartha chose to focus on deep impact by doing more for the same communities in one region.

Thinking critically about funding and timelines. | When evaluating projects to be funded, including pilots and interventions resulting from design-thinking projects, funders typically use criteria that define success as linear growth, while design projects in international development and the social sector evaluate success in terms of people's developmental and behavioral changes, as well as long-term societal changes. In her work with practitioners using design for social innovation, Northumbria University professor of design and social innovation Joyce Yee found that practitioners encounter a disconnect between the evaluation criteria predefined by the funders (based on cost-based models and efficiency and numbers) and the work they do with communities.[12] She advocated for framing evaluations not as strict measures of success but as learning opportunities where the project outcomes could be continuously redefined and cocreated with communities. In practice, this means providing more flexible funding not tied to predetermined outcomes, which would make evaluation reporting less burdensome and allow for more agility and responsiveness to changing circumstances.

The structures of funding in the social sector and international development are not conducive to sustained community engagement. A critical design stance acknowledges the need for trust and relationship building in work with communities. For example, Dalberg Design and Project Concern International India collaborated on two programs to increase men's participation in family planning and nutrition. The projects began in January 2020 and lasted for more than two years. The pilot program, composed of 2,000 households in Bihar, a state in East India, showed potential impact from children's dietary diversity, use of modern contraceptives, and increased involvement of men in usually gendered labor like feeding children. While Dalberg Design's cofounder Robert Fabricant observed that "establishing long-term partnerships with community-based organization is essential," he also noted that "sustaining those partnerships with a small team and limited resources is not easy." Sustained community engagement and trust require extensive resources, which not all nonprofits have.

Practicing Your Stance

CRITIQUES OF DESIGN THINKING ARE not new, and several alternatives, such as systems-design thinking, have been proposed—including in this publication.[13] These alternatives build on

critiques similar to ours but contend that what is needed is a new or augmented methodological tool kit.

Instead, we urge those who practice design thinking to ground their practice in a critical design stance and to be reflexive and deliberate about the intentions, actions, and effects of their work. To begin this work, we invite you to consider the following questions and recommendations:

WHO IS INVOLVED IN THE PROJECT, AND IN WHAT CAPACITY?

- Make sure to define what participation looks like in your work and how communities are involved and engaged in the process.
- Challenge the myth that designers are neutral agents. Acknowledge and address unequal power dynamics and reflect on your positionalities, assumptions, and biases and how they may affect your work and your relationships with communities.
- Be mindful of the toll it takes for communities to participate in the work and deploy trauma-responsive practices to minimize potential harm.
- Seek to improve the material conditions of the communities you work with. Ensure that everyone is fairly compensated for participating in the work by including appropriate compensation in your project budget.

WHO SHOULD FRAME AND SCOPE THE PROJECT?

- Make sure that all stakeholders—especially community members—are involved in determining the frame and scope of the project, and plan for the time needed to ensure their participation.

WHAT OUTCOMES SHOULD BE PRIORITIZED?

- Ensure that you are not embarking on a project that seeks innovation for innovation's sake. Be open to outcomes that may not be novel but are nonetheless effective for communities.

TO WHOM IS THE PROJECT ACCOUNTABLE?

- Research the parties and stakeholders you are working with to understand what motivations they have and how they will benefit from the project. Determine whether the project may legitimize systems that harm the most marginalized.
- Be prepared for the possibility that your project may point in a direction that should not be pursued if it is not aligned with the community's goals. Build accountability measures to prevent outcomes that are not aligned with the community's goals from being considered or implemented.

HOW SHOULD THE PROJECT'S IMPACT BE MEASURED?

- Consider what success looks like for your project, balancing deep impact with market-based conceptions of scale. Revisit your definitions of impact and scale accordingly, and align your decisions with those definitions when determining where to invest your resources.

WHAT FUNDING MECHANISMS AND TIMELINES COULD SUPPORT SUSTAINED IMPACT?

- Consider structuring funding and timelines with more flexible end dates to account for the shifting realities of implementation and allow for agility and responsiveness accordingly.
- Invest in building relationships with communities. Identify and work in solidarity with community-based organizations, and build time for relationship and trust building into your project timeline. Prioritize spending energy and efforts in building longer-term partnerships that outlast the project's life cycle.

To further deepen your critical design stance, we invite you to learn more about and center your work in community-led, decolonial, and other liberatory and anti-oppressive design practices and movements, such as design justice, decolonizing design, and pluriversal design.[14]

We hope that you will continue seeing beyond the prescriptive nature of tools and methods and into the values that can ground your work and your commitment to the communities you serve. ●

ANNE-LAURE FAYARD is the ERA Chair in Social Innovation at NOVA School of Business and Economics and visiting research faculty at New York University (NYU). She is the coordinator of the Design for Social Innovation and Sustainability Lab at NOVA SBE and a founding faculty member of the Design Lab at NYU MakerSpace and Design for America of NYU.

SARAH FATHALLAH is a social designer, researcher, and educator who writes and speaks about the ethics and trauma responsiveness of research and design. They are a senior research fellow at Think of Us and serve on the faculty of the California College of the Arts, UC Berkeley Extension School, and Pratt Institute.

NOTES

1 Sarah Fathallah and A.D. Sean Lewis, "Abolish the Cop Inside Your (Designer's) Head: Unraveling the Links Between Design and Policing," *Design Museum Magazine*, vol. 18, 2021.
2 Lorraine Gamman and Adam Thorpe, "Makeright—Bags of Connection: Teaching Design Thinking and Making in Prison to Help Build Empathic and Resilient Communities," *She Ji: The Journal of Design, Economics, and Innovation*, vol. 4, no. 1, 2018.
3 Jocelyn Wyatt et al., "The Next Chapter in Design for Social Innovation," *Stanford Social Innovation Review*, vol. 19, issue 1, 2021.
4 Anne-Laure Fayard, "Making Time for Social Innovation: How to Interweave Clock-Time and Event-Time in Open Social Innovation to Nurture Idea Generation and Social Impact," *Organization Science*, September 28, 2023.
5 Arturo Escobar, "Designing as a Futural Praxis for the Healing of the Web of Life," in *Design in Crisis: New Worlds, Philosophies and Practices*, edited by Tony Fry and Adam Nocek, London: Routledge, 2020.
6 Sasha Costanza-Chock, *Design Justice: Community-Led Practices to Build the Worlds We Need*, Cambridge, Mass.: MIT Press, 2020.
7 There are several design theorists and practitioners whose work could be named here. For the sake of brevity, we offer three. On sustainability, see Victor Papanek, *Design for the Real World: Human Ecology and Social Change*, New York: Pantheon Books, 1971; on decolonization, see Elizabeth Tunstall, *Decolonizing Design: A Cultural Justice Guidebook*, Cambridge, Mass.: MIT Press, 2023; on anticapitalism, see Matthew Wizinsky, *Design After Capitalism: Transforming Design Today for an Equitable Tomorrow*, Cambridge, Mass.: MIT Press, 2022.
8 Sarah Fathallah, "Why Design Researchers Should Compensate Participants," *Notes Off the Grid*, Medium, April 7, 2020; Sarah Fathallah, "Confronting the Power Designers Wield." *UX Collective*, September 10, 2021; Sarah Fathallah, "Trauma Responsiveness in Participatory Research," Think of Us, September 9, 2022; Sarah Fathallah, "An Ethic of Care for Research Participants as Trauma Survivors," Think of Us, April 3, 2023.
9 Chicago Beyond, *Why Am I Always Being Researched?*, Equity Series, vol. 1, 2018.
10 See Anne-Laure Fayard et al., "Designing Services at Engine (B): Co-designing for Health and the Domestic Environment," Case Reference 411-021-1, the Case Centre, 2011.
11 Sarah Fathallah and Sarah Sullivan, *Away From Home: Youth Experiences of Institutional Placements in Foster Care*, Think of Us, July 21, 2021.
12 Joyce Yee et al., "Measuring Impact," in *Design for Social Innovation: Case Studies from Around the World*, edited by Mariana Amatullo et al., New York: Routledge, 2022.
13 Thomas Both, "Human-Centered, Systems-Minded Design," *Stanford Social Innovation Review*, March 9, 2018.
14 In addition to the references mentioned throughout this article, we offer the following works for exploration: Arturo Escobar, *Designs for the Pluriverse: Radical Interdependence, Autonomy, and the Making of Worlds*, Chapel Hill, N.C.: Duke University Press, 2018; Batya Friedman and David G. Hendry, *Value Sensitive Design: Shaping Technology with Moral Imagination*, Cambridge, Mass.: MIT Press, 2019; Kat Holmes, *Mismatch: How Inclusion Shapes Design*, Cambridge, Mass.: MIT Press, 2020; Claudia Mareis et al., eds., *Design Struggles: Intersecting Histories, Pedagogies, and Perspectives*, Amsterdam: Valiz, 2021; Ron Wakkary, *Things We Could Design for More Than Human-Centered Worlds*, Cambridge, Mass.: MIT Press, 2021; and op. cit. Papanek's *Design for the Real World* and Fry and Nocek's *Design in Crisis*.

Discovery and Validation of Business Models: How B2B Startups can use Business Experiments

Patrick Brecht, Daniel Hendriks, Anja Stroebele,
Carsten H. Hahn, & Ingmar Wolff

> *"The true method of knowledge is experiment."*
>
> William Blake
> Poet, painter and inventor
> "The Argument" (1788)

Startups searching for a business model face uncertainty. This research aims to demonstrates how B2B startups can use business experiments to discover and validate their business model's desirability quickly and cost-effectively. The research study follows a design science approach by focusing on two main steps: build and evaluate. We first created a B2B-Startup Experimentation Framework based on well-known earlier frameworks. After that, we applied the framework to the case of the German startup heliopas.ai. The framework consists of four steps (1) implementation of a measurement system, (2) hypothesis development and prioritization, (3) discovery, and (4) validation. Within its application, we conducted business experiments, including online and offline advertisements, as well as interviews. This research contributes in several ways to the understanding of how B2B-startups can use business experiments to discover and validate their business models: First, the designed B2B-Startup Experimentation Framework can serve as a guideline for company founders. Second, the results were used to improve the existing business model of the German B2B startup heliopas.ai. Finally, applying the framework allowed us to formulate design principles for creating business experiments. The design principles used in the study can be further tested in future studies.

1. Introduction

In recent years, a common approach and conventional wisdom has urged founders to create a business plan that describes the size of the opportunity, the targeted problem, and the planned solution (Blank, 2013). It assumes that the target market is known, and the business model is validated (Garvin, 2000). However, these conditions are often not met by startups, which leads to many startups failing when executing an assumption-based business model (Lynn et al., 1996).

To support founders in searching for a business model, frameworks were created with the idea of conducting experiments in business settings. Technological advances in the last decade have lowered market entry barriers and the cost of running business experiments dramatically (Kerr et al., 2014). This has made business experimentation more viable and simpler to execute. Moreover, characteristics specific to the B2B market influence how business experiments are designed. An implication for a B2B business is that more money is generated by fewer customers (Croll & Yoskovitz, 2013). This implies that it is more difficult to provide statistical significance because of small sample sizes. Additionally, the role of decider and user might not come from the same person as usual with consumers, making it harder for a startup to sell a product to a company (Croll & Yoskovitz, 2013). Thus, more research is needed to determine how B2B startups can specifically use business experiments, allowing founders to learn about the business model. Recently, the COVID-19 pandemic has limited personal contact with customers and thus influenced the methodology of this research. These trends demand additional research in the field.

In line with Berglund, Dimov and Wennberg (2018), who call for more research resulting in practical insights for

Discovery and Validation of Business Models: How B2B Startups can use Business Experiments

Patrick Brecht, Daniel Hendriks, Anja Stroebele, Carsten H. Hahn, & Ingmar Wolff

entrepreneurs, the goal of this research is twofold. First, the study provides empirical insights on the application of business experiments to the business model development process of B2B companies. Second, we investigate how to run and design these experiments. Berglund et al. (2018) recommended creating context-specific design principles in the form of pragmatic recommendations. Thus, this research focuses on extracting practical design principles that support entrepreneurs in improving their business experiment activities. This research takes a problem-solution approach aimed at extracting practical contributions for B2B startups, instead of focusing on enriching the existing theoretical body of literature. The study answers the following question: How can startups in a B2B market use business experiments to discover and validate the desirability of their business models quickly and cost-effectively?

To answer this question, we followed a two-step process proposed by March and Smith (1995), and tailored the research process by focusing on the case of a real life startup named heliopas.ai. The case of heliopas.ai suits this research well as it is searching for a business model that incorporates selling an application called WaterFox to farmers (a B2B context). The mission of the startup heliopas.ai is to provide farmers with accurate data about soil moisture combined with simple recommendations for more efficient field irrigation. The startup uses machine learning and multiple data sources such as satellite imagery and local weather databases to gather the data. No equipment or high-priced sensors are necessary to use the application WaterFox, an advantage deemed beneficial as it saves customers unnecessary costs and adoption efforts. A framework was created that will serve as a guideline to conduct tailored business experiments for heliopas.ai that consider the limitations of startups regarding time and money. These business experiments would help to discover and validate the desirability of their business model in a B2B market. As the startup is based in Karlsruhe, Germany, the initial market consisted of local farmers in Baden-Wuerttemberg and Rhineland-Palatinate.

This research proposes a B2B-Startup Experimentation Framework with a four-step solution that reduces uncertainty and improves a startup's business model. The framework builds on existing processes and principles, and combines them in one comprehensive framework that serves as a guideline for conducting B2B business experiments. By applying the framework to the context of heliopas.ai, the researchers were able to evaluate the proposed framework's applicability. Additionally during the research, the business and operations of heliopas.ai were adjusted due to the findings, resulting in better understanding of the suitability of channels, value proposition, customer jobs-to-be-done, customer segment, and product performance.

This paper is structured as follows. The next section provides a brief review of theories and frameworks used to develop the B2B-Startup Experimentation Framework. In section 3, we lay out the methodology. In the consecutive parts, we develop the framework and apply it to the startup heliopas.ai. Next, we summarize the findings and further elaborate in the discussion section. Finally, we discuss the limitations of the study, practical implications for startups and researchers, and conclude the paper.

2. Theoretical Framework

Discovery and Validation

Discovery marks the initial step in the search for a business model. The goal is to explore if the general direction of thought regarding a business model is correct and to gain more insights (Bland & Osterwalder, 2020). Discovery suits the early steps of experimentation. Since a startup operates under great uncertainty (Ries, 2011), decision making is done under ambiguity, with little to no knowledge about alternatives and consequences (Cooremans, 2012). Generally, validation activities ensure that customers' needs and the defined requirements are met (Albers et al., 2017). The validation process of a business model determines and ensures a correct direction of thought, and also confirms findings from the discovery step (Bland & Osterwalder, 2020). Thus, validation becomes the second step in the search for a business model (Bland & Osterwalder, 2020).

Business Model and Risk Factors

According to Brown and Katz (2009), an early business model entails three risk factors: desirability, feasibility, and viability. Desirability shows the risk of a business model regarding the market, demand, communication, and distribution. Feasibility defines the risk when a business cannot access key resources, perform key activities, or find key partners (Brown & Katz, 2009). Viability denotes the risk that a business cannot generate

Discovery and Validation of Business Models: How B2B Startups can use Business Experiments

Patrick Brecht, Daniel Hendriks, Anja Stroebele, Carsten H. Hahn, & Ingmar Wolff

sufficient revenue or requires too much cost to make a profit; that it won't be viable (Bland & Osterwalder, 2020). This research focuses on reducing the desirability risk in business models. It therefore focuses on the following business model components: customer segments, value proposition, channel, and customer relationship. Additionally, we explore a revenue model in terms of a customer's willingness to pay for heliopas.ai's application offer.

Business Experiments

Business experiments attempt to take a scientific approach for generating insights into a company's business model (Thomke, 2019). They reduce risk and uncertainty by yielding evidence regarding an underlying hypothesis (Bland & Osterwalder, 2020). Experiments can demonstrate a causal relationship by measuring the effect an action has on a situation (Hanington & Martin, 2012). Business experiments in startups are often run cheaply and quickly (Aulet, 2013). For this paper, business experiments can be distinguished based on how they are aligned with their purpose — discovery and validation — as in the previously provided definitions. Discovery experiments test if the general idea behind a concept is acceptable, intending to establish a proof-of-concept. Validation experiments are experiments with higher fidelity. They yield stronger evidence and require more resources, such as time, personnel or money.

Growth Hacking

Growth Hacking aims at fast and sustainable growth through activities in the area of market research, product development, and customer retention (Ellis, 2017). The Customer Acquisition Funnel is a core element of the Growth Hacking framework. It consists of five stages: acquisition, activation, retention, revenue, and referral (McClure, 2007). In the acquisition stage, the goal is to figure out through which channels users, customers, and visitors are coming from, in a way that results in value for a startup (McClure, 2007). Secondly, the activation element shows how many acquired users have a positive first impression of the product (McClure, 2007). Retention measures whether users keep using the product. The revenue stage measures customers' willingness to pay, whereas the referral stage measures if users enjoy the product enough to recommend it to a friend (McClure, 2007).

Customer Development Process

The Customer Development Process by Blank and Dorf (2012) is an iterative, customer-focused approach in the search for a business model. It incorporates business experiments and consists of two steps: customer discovery and customer validation. Customer discovery aims at finding initial customers by deriving testable hypotheses that collect possible experiment designs (Blank & Dorf, 2012), and finally conducting the experiments. These initial experiments determine if the envisioned value proposition matches a targeted customer segment. In the next step, the proposed solution is presented to customers to learn if it serves customer needs, and to assess customer willingness to pay for it. Customer validation requires applying the business model that results from the previous step (Trimi & Berbegal-Mirabent, 2012). The goal is to test whether the business model is repeatable and scalable. This is done by running more quantitative, high-fidelity experiments and acquiring actual sales, which will show how money spent in sales and marketing can generate revenue.

Lean Startup

The Lean Startup aims to reduce waste while creating a business model. It has three key principles: to replace planning with experimentation, the 'getting out of the building' approach by Blank, and lastly, agile development (Blank, 2013). The experimentation process is described by the Build-Measure-Learn feedback loop consisting of three steps: build, measure, and learn. In the step build, it is essential to create a Minimal Viable Product (MVP) quickly after identifying the most important hypotheses (Ries, 2011). The goal of building an MVP is to identify the proposed solution's potential (Kerr et al., 2014) and the target customers' willingness to pay for it. The measure step aims at collecting data that can verify or falsify the hypotheses made about product quality, price, and costs (Ries, 2011). In the learn step, the goal is to learn about the investigated hypotheses from collected data. The learning process shows whether an underlying hypotheses can be verified or not, and indicates if the MVP is a viable solution to the customer problem (Ries, 2011).

Four-Step Iterative Cycle

The Four-Step Iterative Cycle describes a structured procedure of business experimentation that undergoes

Discovery and Validation of Business Models: How B2B Startups can use Business Experiments

Patrick Brecht, Daniel Hendriks, Anja Stroebele, Carsten H. Hahn, & Ingmar Wolff

Table 1. Framework comparison

(a) Comparison of Micro Cycle

	Pre-Experiment			Experiment		Post-Experiment
Growth Hacking	Ideate	Prioritize		Test		Analyze
BML* Feedback Loop			Build	Measure		Learn
Four-Step Iterative Cycle		Design	Build	Run		Analyze
B-SEF** Micro Cycle	Ideate and Design		Build	Run and Measure		Analyze and Decide

(b) Comparison of Macro Cycles

		Customer Discovery				Customer Validation			
Customer Development Process		Hypothesis Collection	Problem Testing	Solution Testing	Verification or Pivot	Sales Preparation	Sale	Sales Refinement	Verification or Pivot
Growth Hacking	Measurement System	Experimentation							
B-SEF** Macro Cycle	Measurement System	Hypothesis Collection	Discovery Experiments			Validation Experiments			

* BML = Build Measure Learn. ** B-SEF = B2B-Startup Experimentation Framework.

the design, build, run, and analyze steps iteratively until it achieves desired outcomes. The first step design uses existing insights from observations and previous experiments to formulate testable hypotheses, and design suitable experiments (Thomke, 2003). In the build step, researchers build physical or virtual prototypes or models to conduct experiments (Thomke, 2003). The higher a prototype's fidelity and functionality, the stronger the generated evidence will be (Thomke, 2003). Subsequently, the experiment is run either in a more controlled laboratory setting or in a real-life setting, which produces higher external validity (Thomke, 2003). Finally, the results are analyzed by comparing them to an expected outcome. If the hypothesis addressed by the experiment is answered sufficiently, the experimentation cycle is stopped (Thomke, 2003). Otherwise, researchers reenter the design step with a modified experimental design, adjusted according to new insights gained in the process. Table 1 summarizes the presented frameworks and shows an initial comparison with the framework designed for this research.

3. Methodology

To create a business experimentation framework for B2B startups and gain insights on how B2B startups can use business experiments to discover and validate their business model quickly and efficiently, this research applied a two-step research process based on "design science" insights, as suggested by March and Smith (1995). The two-step research process in design science consists of a build and evaluate step, which can be summarized as follows.

The build step for this paper undertakes a literature review to shape a framework based on existing knowledge and practical experiences of respected practitioners (Thomke, 2003; Ries, 2011; Blank, 2012; Ellis, 2017), as well as previous research conducted in this field (Thomke, 2003). Practical knowledge is very popular among entrepreneurs. Although it is not grounded in theory itself, it is considered a valuable source of knowledge in this field.

Discovery and Validation of Business Models: How B2B Startups can use Business Experiments

Patrick Brecht, Daniel Hendriks, Anja Stroebele, Carsten H. Hahn, & Ingmar Wolff

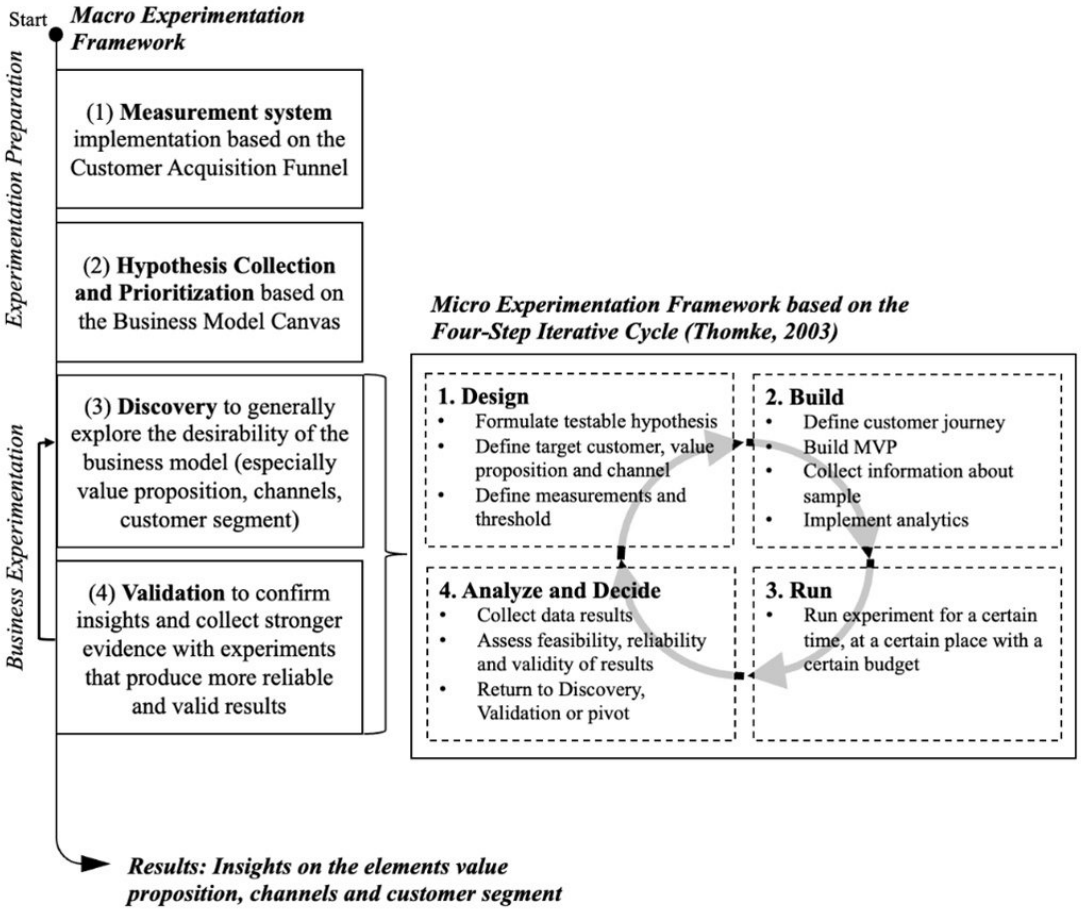


Figure 1. The B2B Startup Experimentation Framework (B-SEF)

Next, we apply the framework to the particular case of heliopas.ai, a real life startup that wants to improve its business model. This constitutes the evaluate step of the two-step process, which aims to show whether the created framework fulfills its purpose. Furthermore, the application allows researchers to deepen their knowledge about how to run business experiments empirically. Qualitative and quantitative data was collected during several business experiments. We used this empirical data to develop insights into heliopas.ai's business model. Also, we describe applying the framework and conducting business experiments, which resulted in formulating design principles that serve as recommendations for conducting business experiments. The design principles can be regarded as a basis for future research that focuses on further investigating the value of business experiments.

4. Design and Application of the B2B-Startup Experimentation Framework

Based on the frameworks described in the theoretical part of this paper, we designed the B2B-Startup Experimentation Framework (B-SEF) and outline it in the following way. It consists of both a macro experimentation and micro experimentation framework (see Figure 1).

The macro experimentation framework consists of four steps. First, it involves designing a simple measurement system to collect data on acquiring and retaining new customers. The idea of implementing such a measurement system originates from the Growth Hacking methodology. Applying it for this research was feasible because the use case startup already has a vision for its business model and technology integrated into a

Discovery and Validation of Business Models: How B2B Startups can use Business Experiments

Patrick Brecht, Daniel Hendriks, Anja Stroebele, Carsten H. Hahn, & Ingmar Wolff

Table 2. Tracked metrics in heliopas.ai's measurement system

Stage	Metric	Explanation
Acquisition	Total Traffic	Amount of Visitors on the Facebook and Web Landing-page of the WaterFox Application
Activation	App store Product Site Impressions	Amount of visitors that opened the app store product site in the Google Play Store and Apple App Store
	App downloads	Amount of app downloads from the Google Play Store and Apple App Store
	New User Registrations	Amount of new users registered in the WaterFox application
Retention	Occasional Users	Users that opened the WaterFox application 1-2 days in the past seven days
	Standard Users	Users that opened the WaterFox application 3-5 days in the past seven days
	Heavy Users	Users that opened the WaterFox application 6-7 days in the past seven days
	Observed Hectares	Amount of hectares that were observed by all registered farmers
Revenue	Paying Users	Amount of users that paid for using the WaterFox application

smartphone-application tested by selected customers. The data is used to calculate conversion rates and customer acquisition costs (CAC), as well as estimate customer lifetime value (CLV). Second, the Business Model Canvas (Osterwalder et al., 2010) is used to collect and prioritize hypotheses about the business.

Business experiments are conducted in the two steps discovery and validation (Bland & Osterwalder, 2020), thereby incorporating, specifically, discovery and validation experiments. By doing so, this research follows the recommendation of Blank and Dorf (2012) who suggest to treat the search for a business model as a two-step process of discovery and validation. In the discovery step, business researchers aim at gaining insights quickly and cost-effectively, as timing can be critical for a startup's success. As emphasized by Ries (2011), the goal is to learn quickly about the business model's desirability. In the validation step, researchers design experiments to gather more reliable evidence. By adding a control group and running experiments simultaneously, the effects of external variables will be reduced.

The micro experimentation framework is adapted from the Four-Step Iterative Cycle by Thomke (2003). All business experiments are conducted and presented in a structured manner by following a micro experimentation framework. The Four-Step Iterative Cycle is a core element of many frameworks and methods for startup experimentation. For instance, Brecht and his colleagues (2020) used the Four-Step Iterative Cycle to conduct experiments for platform business models. Ries (2011) and Blank and Dorf (2012) described similar cycles for experimentation in their Lean Startup and Customer Development Process.

The B-SEF was applied to the startup heliopas.ai to gain insights into its business model and to empirically evaluate the framework's applicability. We note as important that restrictions of cost and time were present in this study, based on a budget of less than €100, and less than four weeks to design and run each experiment.

(1) Designing the Measurement System

Growth Hacking relies on experiments, and thus must collect data. A common way to determine how to design a measurement system that suits the purpose of data collection is to use the Customer Acquisition Funnel, described above. Consecutively, we designed the customer journey for potential customers of heliopas.ai based on multiple metrics, which were defined and tracked. Table 1 provides an overview of these metrics, their definition, and their application in the Customer Acquisition Funnel stages.

Discovery and Validation of Business Models: How B2B Startups can use Business Experiments

Patrick Brecht, Daniel Hendriks, Anja Stroebele, Carsten H. Hahn, & Ingmar Wolff

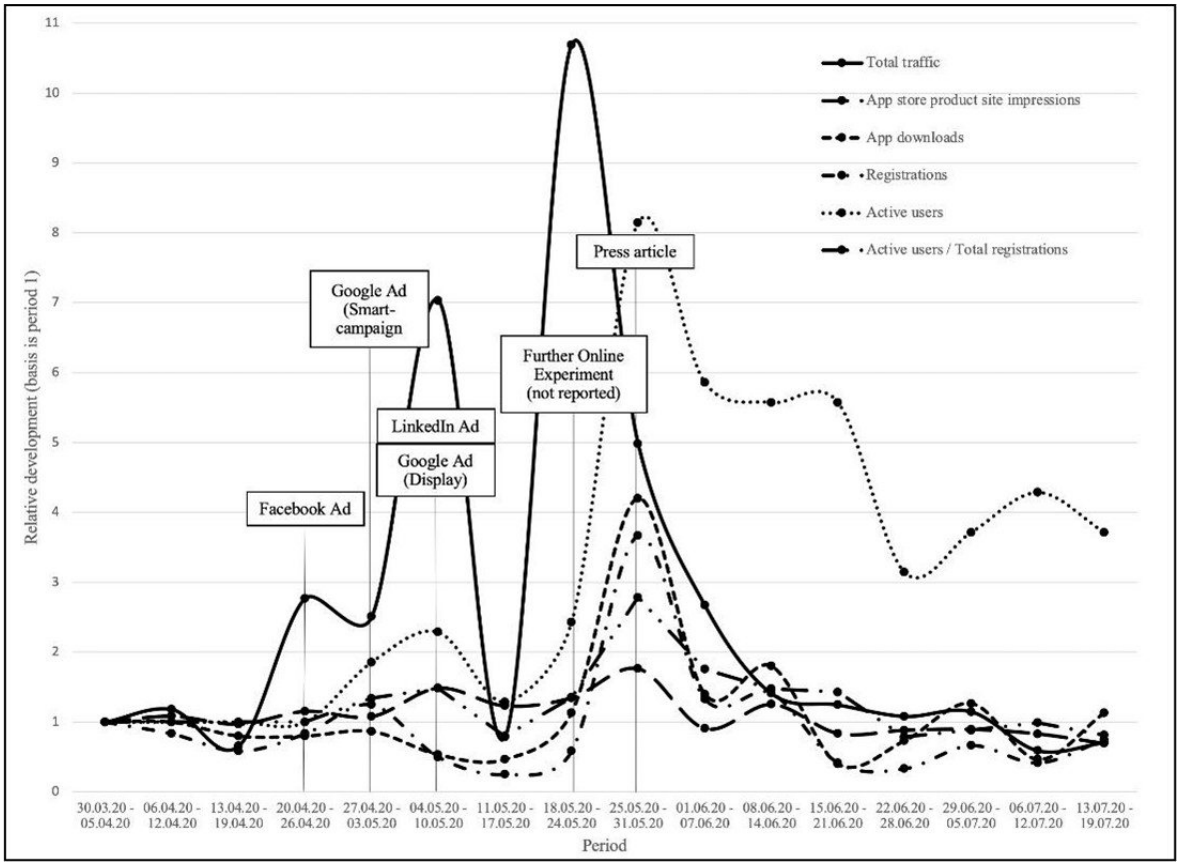


Figure 2. Development of Metrics of the Customer Acquisition Funnel

The acquisition stage consisted of metrics on total traffic generated by sites related to WaterFox. We extracted traffic data on the Facebook landing-page from Facebook's analytics. We tracked traffic on the WaterFox web landing-page with Google's analytics.

The activation stage consisted of metrics from app store product site impressions in the Apple and Google Play app stores. Additionally, we tracked the downloads of the WaterFox application from both app stores and new user registrations in the WaterFox application. Likewise, App store product site impressions and downloads were tracked with the App Store Connect and Google Play Console.

For the retention stage, we defined three metrics. The users were split into the three categories, occasional user, standard user, and heavy user, based on frequency of signing into the application in the last seven days. Since small errors in data had a high impact (for example, activities of developers in the application inflating the data), it was necessary to get data first-hand that was adaptable and transparent. The data from the retention stage was measured by building an Excel sheet that processed data from the startup's database. This data was used to calculate daily retention metrics. Beneficial to this approach was that the researchers could manually filter users, since user names and further contact information were available.

The revenue stage consisted of metrics from paying users, who were paying to use the WaterFox application.

The referral stage was not tracked due to focus on the other stages. The collected data is presented in Figure 2. The three user categories are summarized as active users.

We used the absolute number metrics for a certain period to calculate the conversion rate between customer journey phases via app store product site impressions to app downloads and app downloads to registrations. We used the conversion rate to estimate Customer Acquisition Costs (CAC), since the customer journey could not be tracked after customers leave the landing page and are referred to the app store.

Discovery and Validation of Business Models: How B2B Startups can use Business Experiments

Patrick Brecht, Daniel Hendriks, Anja Stroebele, Carsten H. Hahn, & Ingmar Wolff

(2) Hypothesis Collection and Prioritization
To collect and prioritize initial hypotheses, we used the common Business Model Canvas for heliopas.ai. Our focus on coming up with a desirable business model, drew upon the building blocks value proposition, customer segment, channels, customer relationships, and revenue model with greatest interest. We prioritized our hypothesis resulting in a focus on channels according to the founders' vision of selling their application online. This would allow them to distribute the app efficiently at a low CAC and easily reach early adopters. Hence, in the following section, our attention will turn to experiments exploring the channels.

(3) Discovery Experiments
For each discovery experiment, we tracked the number of impressions, clicks on the advertisement leading to the landing page, and download-button clicks on the landing page to evaluate channel suitability. We calculated the Customer Activation Rate (CAR) from recorded data. CAR is defined as the number of active persons on the landing page, divided by the number of visitors. A threshold of 20%, a common value for experiments in a startup environment, was defined for CAR. Additionally, the calculated CAC is going to help to evaluate the channel's viability. Table 2 provides an overview of the discovery experiments conducted.

Facebook advertisement experiment. The Facebook advertisement experiment analyzed whether customers were pulled to the WaterFox application via advertisements on Facebook. In the Facebook advertisement manager, the target customer was set to the current persona. A customer journey was designed, leading customers from an advertisement on landing page, to the app store, and finally to the application. To measure all online activities on the landing page, we

Table 3. Overview of conducted Discovery Experiments

Experiment	Facebook	Google (Smart-Campaign)
Hypothesis	*Vegetable and potato farmers in Germany, Austria and Switzerland can be pulled to the WaterFox application via advertisements on Facebook.*	*Vegetable and potato farmers in Germany, Austria and Switzerland can be pulled to the WaterFox application via advertisements on Google Ads.*
Test	Run Advertisement in Facebook	Run Google Smart-Campaign
Metric	Impressions, Clicks and Downloads	Impressions, Clicks and Downloads
Criteria	$CAR \geq 20\%$	$CAR \geq 20\%$
Run Time	7 Days	3 Days
Budget	€31.20	€27.31

Experiment	Google (Display)	LinkedIn
Hypothesis	*Vegetable and potato farmers in Germany, Austria and Switzerland can be pulled to the WaterFox application via advertisements in the Google Display Network.*	*Vegetable and potato farmers in Germany, Austria and Switzerland can be pulled to the WaterFox application via advertisements on LinkedIn.*
Test	Run Google Display Advertisement	Run Advertisement in LinkedIn
Metric	Impressions, Clicks and Downloads	Impressions, Clicks and Downloads
Criteria	$CAR \geq 20\%$	$CAR \geq 20\%$
Run Time	2 Days	6 Days
Budget	€29.98	€30.00

Experiment	Press Article
Hypothesis	*Farmers in Germany, Austria and Switzerland can be pulled to the WaterFox application by an article in the agriculture newspaper top agrar.*
Test	Release Press Article
Metric	Impressions, Clicks and Downloads
Criteria	$CAR \geq 20\%$
Run Time	7 Days[9]
Budget	€0

Discovery and Validation of Business Models: How B2B Startups can use Business Experiments

Patrick Brecht, Daniel Hendriks, Anja Stroebele, Carsten H. Hahn, & Ingmar Wolff

implemented Google Analytics and Google Tag Manager. The advertisement was run for seven days from April 22 to April 28, 2020, with a budget of €31.20 resulting in 4.557 impressions, 5 clicks on the advertisement, and 0 download-button clicks on the landing-page leading to a CAR of 0%, which was below the set threshold of 20%. Therefore, the hypothesis was falsified, and we decided to try a different channel in the next discovery experiment.

Google advertisement experiments. The Google advertisement experiment investigated whether customers are pulled to the WaterFox application via Google Ads. It ran as a smart campaign, which means that bidding, targeting, and ad creation were automated by Google Ads (Google, 2020). The landing-page from the Facebook advertisement experiment was reused, providing customers with an almost identical customer journey. The experiment ran for three days from April 27 to April 29, 2020, with a budget of €27.31, resulting in 48.211 impressions, 89 advertisement clicks, and 4 download-button clicks. This resulted in a CAR of 4.49%, which is also lower than the expected 20%. Hence the hypothesis was falsified. The Google Ads manager provides further information about the keyword performance, types of devices by targeted customers, and advertisement networks. The highest CAR in the Google Display Network was 11.11%. Due to this performance indication, we decided to conduct a second Google advertisement in the Google Display Network.

The Google Display Network experiment used a display advertisement that was only shown on certain websites and not in Google Search. The experiment ran for three days from May 6 to May 7, 2020, with a budget of €29.98, resulting in 69.100 impressions, 539 advertisement clicks, and 8 download-button clicks. This led to a CAR of 0.15%, which was below the set threshold of 20%. Therefore, the hypothesis was falsified. Due to this result, we decided to explore other online channels.

LinkedIn advertisement experiment. The LinkedIn advertisement experiment investigated whether customers were pulled to the WaterFox application via advertisements on LinkedIn. Hence, a LinkedIn advertisement was designed to test the hypothesis. The potential customers were sent to a landing page that revealed detailed information about the value proposition of the WaterFox-application. The target audience was defined as persons interested in agricultural topics. The target audience was set to males between the age of 25-34 years, meant to represent the startup's current target customer persona. For this experiment, the customer journey from the previous experiment was again reused, with the only difference that now the customer started at the designed LinkedIn advertisement. Data was collected using LinkedIn's campaign manager connected to the landing page by implementing a JavaScript tag to count conversions on the landing page. The LinkedIn experiment ran from May 5 until May 10, 2020, with a budget of 30 in Germany, Austria, and Switzerland, resulting in 6 clicks on the advertisement, and 2 download-button clicks. The results meet the expectation threshold with a CAR of 33.33%.

Press article experiment. The press article experiment tested whether customers could be pulled via an article in an agricultural newspaper into the WaterFox application. The hypothesis was formed by interviews that the startup conducted with customers during the business experimentation process. To test the hypothesis, we sent a press release to several newspapers. To contact them, we used the network of the local startup accelerator for support. The press release contained important information about the WaterFox-application. Underneath the article, a link referred directly to the landing-page of the WaterFox application. Referrals from this link were tracked using Google Analytics. The newspaper top agar

Table 3. Results of Discovery Experiments.

Experiment	CTR	CAR	CAC
Facebook Advertisement	0.11%	0.00%	-
Google Advertisement (Smart-Campaign)	0.18%	4.49%	€24.39
Google Advertisement (Display)	0.78%	0.15%	€38.36
LinkedIn Advertisement	0.96%	33.33%	€153.85
Press Article	18.80%	27.47%	€15.93

(https://www.topagrar.com) published the article on May 27, 2020. The press article resulted in 484 unique article reads, 91 landing-page visitors, and 25 download-button clicks. This leads to a CAR of 27.47%, which therefore meets our threshold expectations.

(4) Validation Experiments
After identifying suitable channels for the startup to acquire new users, our focus shifted from the discovery phase to the validation phase. In the following, we present two validation experiments, called brochure advertisement and validation interview.

Brochure experiment. The brochure advertisement experiment tested whether customers were willing to pay a price of €9 for the envisioned version of the WaterFox application and whether customers can be acquired via post. Again, a customer journey was set up. To test customer willingness to pay, two versions of the brochure were designed that differed in price. The control group received a brochure costing €3, while the test group received a €9 brochure. If the post-delivery channel was to be a suitable way of acquiring customers, it would be validated or refuted by running the validation interview experiment afterward.

To run this experiment, we needed the addresses of farmers to send the brochures. To solve this, we screened several websites and platforms for contacts. The sample size was 34, equally divided between test and control group. We sent the brochures to recipients at a cost of €47.73 for printing and sending. The brochure advertisement resulted in no responses from the contacted persons. Follow-up validation interviews with five contacted farmers (referred to as interview experiment in the following), revealed a possible reason for non-responses and disclosed valuable information for the business model.

Interview experiment. Besides trying to discover the cause for the non-responses to the brochures from potential customers, the goal of the interview experiment was to validate current understanding of problems and customer work involving farm irrigation management. More precisely, the experiment aimed to investigate the importance of certain tasks to target customers. Importance was defined as the frequency and effort of completing a task or enduring a burden. Additionally, our goal was to investigate the current usage of digital products for target customers. Job inquiries were formulated as statements. Each statement included whether the customer actually worked in the job, how often it was completed in the last four weeks, and how much time and money were required. Additionally, interviewees were queried about several elements of their farm. This was done to place the provided information into the right context and to help avoid biased or misleading results. The interview was conducted via phone. We read out the statements about jobs to the participants and recorded their responses. Of the 34 contacts available from the previous brochure experiment, we contacted 31. Five contacts agreed to an interview. The remaining showed an unwillingness to be interviewed, mostly due to time pressure and a high workload, because seasonal workers were limited due to the COVID-19 pandemic. The interviews were encoded to categorize the answers and systematically extract the results. We present the results from the validation experiments briefly in the following section.

5. Results

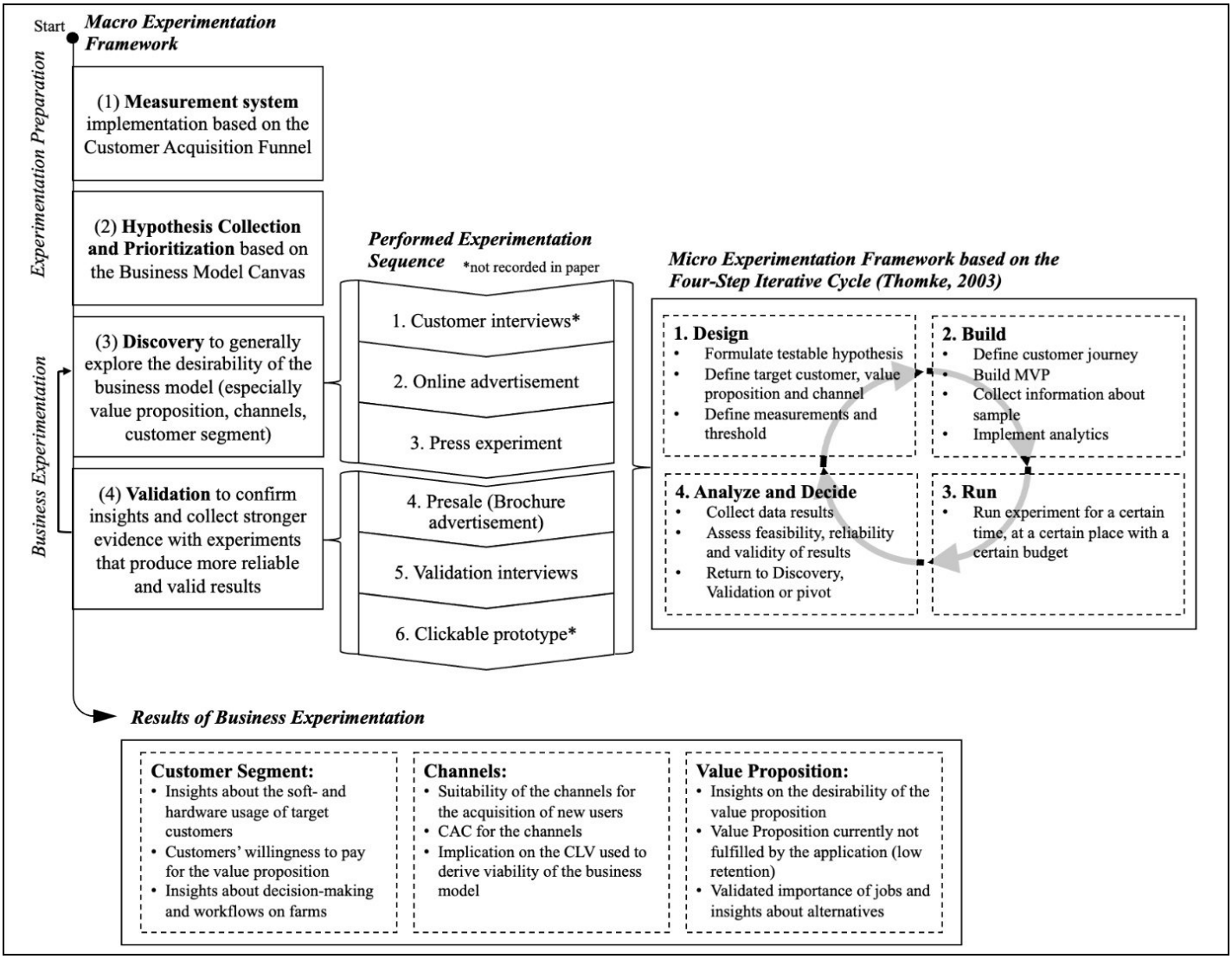
We gained insights into the company's channels, value proposition, customer segment, and product performance. Table 3 summarizes the experimental results conducted in the discovery phase of the B-SEF. The collected data shows the Facebook and Google Ads did not reach the predetermined conversion threshold of 20%. In contrast, the LinkedIn advertisement and press article experiment exceeded the threshold. The cost of acquiring one registered user was €153.85 for LinkedIn and €4.13 for the press, based on the measurement system and data collected by running the experiments. We estimated the cost of running the press article experiment based on the average price of using a writing service from a freelancer on the website upwork.com. These results were valuable for the startup to evaluate the desirability of its business model as they showed how the startup can acquire new customers and how much it costs.

The distributed brochures did not result in any acquisitions. The follow-up interviews conducted to investigate the unresponsiveness revealed that customers have limited available time and chose not to allocate it to reading a brochure. Additionally, the interviews yielded insights into software and hardware usage, as well as willingness to pay for the product. These insights helped to evaluate how desirable certain elements of the business model were, such as the value proposition. Figure 3 summarizes key findings of the B-

Discovery and Validation of Business Models: How B2B Startups can use Business Experiments

Patrick Brecht, Daniel Hendriks, Anja Stroebele, Carsten H. Hahn, & Ingmar Wolff

Figure 3. The B2B Startup Experimentation Framework (B-SEF) with Results



SEF application in the startup heliopas.ai.

6. Discussion of Results and Proposed Design Principles

This research set out to assess the business model of heliopas.ai. By conducting a series of business experiments outlined by the B2B Startup Experimentation Framework (B-SEF) with an emphasis on the discovery and validation step of a startup, valuable insights were gained, resulting in improvements to the business model. With only a low budget of €118.49, experimentation showed the desirability of the business model and revealed the presented results which are discussed in the following.

The discovery experiments were run at various times, which led to extraneous variables not remaining constant. This was an acceptable circumstance of the discovery experiments as they were aimed at establishing proof of the channel's suitability in reaching target customers quickly. With a CAR of 27.47%, the press article experiment met our threshold expectations. A possible reason for this performance might be that farmers trust the information delivered by the newspaper top agrar, and are therefore more likely to visit the landing page and download the WaterFox application. If increased trust leads to more landing page

visits and conversions, this article can also be used in future experiments as a reference.

The results of the brochure advertisement emphasize the importance of an existing channel to a customer. The post-delivery channel was not validated and hence the brochure experiment does not provide answers about customer willingness to pay. The validation interview showed that interviewees use different hardware and software, and the difficulty of integrating the WaterFox application into an existing customer workflow. One interviewee stated that they were annoyed by documenting information in several IT systems. Thus, it was an obstacle for helipas.ai to sell its product to other businesses and find innovators.

The result of the interview experiment reveals the benefit of a qualitative approach. This experiment yielded significant insights into customers' jobs, the suitability of the posting channel (used for the brochures), as well as customer willingness to pay. Since all interviewees were tested under all conditions, this was a within-subjects experiment (Price et al., 2017). It provided a high level of control over the extraneous variables, since participants in all control and test groups were the same. This was an advantage of the validation interview experiment's design.

The novelty of the designed framework is that it was tailored to a real B2B startup. Though this in some ways might limit its generalizability, at the same time it also increases its suitability for this particular case of a startup company. By advising the process to begin with implementing a measurement system as the first step, this research stands in contrast to the business experimentation frameworks of Blank and Dorf (2012), Ries (2011), and Thomke (2003). Our measurement system was adapted from the Growth Hacking framework. With heliopas.ai, it was justifiable to build a measurement system at first since the stage and the progress the founders were at with the startup was more advanced at the point of time of this research. Similar to

Table 5. Proposed Design Principles

Case-specific circumstances: B2B startup, high-fidelity MVP existing, software-based product, technology for the application has been developed		
Intention	**Action**	**Recommendation**
Testing online communication and acquisition channels.	Google Ads advertisements LinkedIn advertisement Facebook advertisement Press article	*H1: Online advertisements are less suitable to learn qualitatively about the customer.* *H2: Increase the run-time to one week to improve reliability.* *H3: B2B customers trust the information delivered by the newspaper* top agrar *and are more likely to visit the landing page.* *H4: Consider Customer Acquisition Rate as well as Customer Acquisition Cost as a threshold for experiments.*
Testing the customers' willingness to pay.	Brochure advertisement	*H5: Test the channel to confirm its suitability before running validation experiments.*
Measuring data along the customer journey from a virtual touchpoint to a smartphone application.	Tracking traffic, conversions, app store views, app downloads, registrations, and user retention	*H6: Tools such as Google analytics and Excel spreadsheets enable data collection and derivation of decisions in the startup.*
Exploring the customers' pains.	Customer interview	*H7: Qualitative interviews should be considered for early exploration of the customers' pains.*

research done by Brecht and colleagues (Brecht et al., 2019) that focused on business experiments for platform business models, the B-SEF focuses on B2B market business models. Compared to other frameworks, the B-SEF suggests concrete experiments. This positions our research uniquely among existing frameworks.

Even with restricted generalizability, these business experiments and the design principles derived can provide other startup founders with ideas about designing their own experiments (see table 4). The design principles from this research were formulated using the following structure: "to achieve X in situation Y, something like Z will help" (Berglund et al., 2018). These design principles are currently at hypotheses stage based on applying the framework and require further empirical research to confirm and validate (or refute) them statistically.

In contrast to other frameworks, this research also provided references on the performance of certain experiments, that is, practitioners can compare quantitative results of this research with their results and evaluate the findings. This comparison can be useful in practice when it is not always clear if an experiment has produced satisfactory results.

The separation of business experiments into discovery and validation experiments was proposed to equally satisfy scientific rigour as well as the entrepreneurial desire for speed and efficiency. This mindset was inspired by the Customer Development Process of Blank and Dorf (2012) and by recently published work on business experimentation by Bland and Osterwalder (2020).

As stated previously, businesses tend to be largely regulated and rational in their buying decisions. It is difficult to evaluate what impact customer willingness to adapt to a new product had, given that the customers in this case were other businesses instead of consumers. Based on experience, target customers (farmers) are only open to new products if value is delivered immediately. This makes it hard for startups to penetrate markets with an unfinished product. The limited number of customers in a B2B market such as the agriculture industry, in which the startup heliopas.ai operates, can be a challenge when conducting business experiments. Having available business contacts can be beneficial when running business experiments, which was seen in the validation experiments.

This study chose a threshold of 20% CAR for discovery experiments, a commonly used threshold to determine the success of an experiment in startup environments. However, we recommend relying not only on one metric, but also on monetary metrics, such as CAC or CLV, to assess the success or failure of an experiment.

7. Conclusion

This research proposed the B2B-Startup Experimentation Framework, a four-step solution for how startups can reduce uncertainty and improve their business model. The framework was tailored to the B2B startup case of heliopas.ai. The main contribution of this research lies in having applied a theoretical-based framework to extract insights into the applicability of the proposed framework. The B-SEF guides a B2B startup in how to conduct business experiments. The startup heliopas.ai gained important insights into customer segment, the value proposition, and its business model channel, and reduced uncertainty by following it. However, business experiments might not always be feasible, since requirements might not be available to properly execute them (see, channel for brochure advertisement).

A limitation of the current B-SEF framework is the focus on desirability aspects of the business model during its application. In contrast, other frameworks like the Customer Development Process (Blank & Dorf, 2012) have a process designed to explore and validate the entire business model. The focus on desirability might limit researchers' capabilities of evaluating the framework holistically, and therefore requires further research. It remains an open question whether the B-SEF is only applicable to B2B startups, which likewise leaves room for future research. We recommend applying the framework in a B2C startup to investigate its applicability in that market.

Although online advertisement is a quick and cost-effective way to gain insights into the desirability of the business model, large amounts of data provided by online advertisement tools helped to conclude causality. Limitations arose when investigating why certain events or results occurred. The anonymity of persons was challenging in this case as we were not able to contact people for further questioning. For instance, the Google Display advertisement did not perform as expected and the plausible explanation was merely based on

Discovery and Validation of Business Models: How B2B Startups can use Business Experiments

Patrick Brecht, Daniel Hendriks, Anja Stroebele, Carsten H. Hahn, & Ingmar Wolff

assumptions.

Another limitation of this research is that only a narrow understanding of the causal relationship between variables could be gained. Also, the current pandemic poses another extraneous effect. The lack of seasonal workers might have influenced the amount of time farmers spent online and on social media activities. Therefore, running the same experiment at a different time of the year or in a different year might yield different results. This underlines the limitations of business experiments in general, which are usually run with a low budget and in a short period of time. Moreover, it is ambiguous if an increased budget or run-time would have led to similar results. Since the conducted discovery experiments focused on online experimentation, the framework is expectedly limited to businesses that can acquire customers and distribute their products online.

As a conclusion remark, we find that the main challenge is to design business experiments in a way that reveals underlying causality. This can be very challenging in a startup where the business and its operations are not yet defined. Furthermore, operating quickly and cost-effectively implies making trade-offs between the reliability and validity of results. Practitioners should consider a sequence of business experiments that are run to improve the company's learning effect, to better explain negative outcomes, and to use a mixed data collection approach. William Blake stated (1788) that experimentation is "the true method" of gaining insights. This also seems to hold true for business model validation in B2B startups. A systematic experimentation framework along with well-designed business experiments can reduce the need for resources such as time and money and help deal with uncertainty and risks.

References

Albers, A., Behrendt, M., Klingler, S., Reiß, N., & Bursac, N. 2017. Agile product engineering through continuous validation in PGE - Product generation engineering. *Design Science*, 3(5): 19. DOI: https://doi.org/10.1017/dsj.2017.5.

Aulet, B. 2013. *Disciplined Entrepreneurship: 24 steps to a successful startup.* Hoboken, New Jersey: Wiley.

Berglund, H., Dimov, D., & Wennberg, K. 2018. Beyond Bridging Rigor and Relevance: The three-body problem in entrepreneurship. *Journal of Business Venturing Insights*, 9: 87-91. DOI: https://doi.org/10.1016/j.jbvi.2018.02.001

Bland, D.J., & Osterwalder, A. 2020. *Testing Business Ideas.* Hoboken, New Jersey: John Wiley & Sons, Inc.

Blank, S., & Dorf, B. 2012. *The Startup Owner's Manual: The Step-by-Step Guide for Building a Great Company.* Pescadero, CA: K & S Ranch.

Blank, S. 2013. Why the lean start-up changes everything. *Harvard Business Review.*

Brecht, P., Niever, M., Kerres, R., & Hahn, C. 2020. Smart platform experiment cycle (SPEC): A validation process for digital platforms. Conference: *Thirteenth International Tools and Methods of Competitive Engineering Symposium* (TMCE 2020).

Brown, T., & Katz, K. 2009. *Change by Design: How design thinking can transform organizations and inspire innovation* (1st ed.), Harper Collins.

Cooremans, C. 2012. Investment in energy efficiency: Do the characteristics of investments matter? *Energy Efficiency*, 5: 497-518. DOI: https://doi.org/10.1007/s12053-012-9154-x.

Croll, A., & Yoskovitz, B. 2013. *Lean Analytics: Use data to build a better startup faster.* Sebastopol, CA: O'Reilly.

Ellis, S. 2017. *Hacking Growth: How today's fastest-growing companies drive breakout success.* New York: Crown Business.

Garvin, D.A. 2000. *Learning in Action: A guide to putting the learning organization to work.* Boston, MA: Harvard Business School Press.

Google. 2020. *About smart display campaigns - Google ads help.* Retrieved July 30, 2020. Accessed online: https://support.google.com/google-ads/answer/7020281?hl=en.

Hanington, B., & Martin, B. 2012. *Universal Methods of Design: 100 ways to research complex problems, develop innovative ideas, and design effective solutions.* Beverly, MA: Rockport Publishers.

Kerr, W.R., Nanda, R., & Rhodes-Kropf, M. 2014. Entrepreneurship as experimentation. *Journal of Economic Perspectives*, 28(3): 25-48. DOI: 10.1257/jep.28.3.25.

Lynn, G.S., Morone, J.G., & Paulson, A.S. 1996. Marketing and Discontinuous Innovation: The Probe and Learn Process. *California Management Review*, 38: 8-37. DOI: https://doi.org/10.2307/41165841.

March, S.T., & Smith, G.F. 1995. Design and natural science research on information technology. *Decision Support Systems,* 15(4): 251-266. DOI: https://doi.org/10.1016/0167-9236(94)00041-2

McClure, D. 2007. *Product Marketing for Pirates.* Retrieved August 12, 2020: https://500hats.typepad.com/500blogs/2007/06/internet-market.html.

Discovery and Validation of Business Models: How B2B Startups can use Business Experiments

Patrick Brecht, Daniel Hendriks, Anja Stroebele, Carsten H. Hahn, & Ingmar Wolff

Osterwalder, A., Pigneur, Y., & Clark, T. 2010. *Business Model Generation: A handbook for visionaries, game changers, and challengers.* Hoboken, NJ: Wiley.

Price, P., Jhangiani, R., Chiang, I.-C., Leighton, D., & Cuttler, C. 2017. *Research Methods in Psychology* (3rd ed.).
DOI: https://dx.doi.org/10.17605/OSF.IO/HF7DQ

Ries, E. 2011. *The Lean Startup: How today's entrepreneurs use continuous innovation to create radically successful businesses.* New York: Crown Business.

Thomke, S.H. 2003. *Experimentation Matters: Unlocking the potential of new technologies for innovation.* Boston, MA: Harvard Business School Press.

Thomke, S.H. 2019. *Experimentation Works: The surprising power of business experiments.* Boston, MA: Harvard Business Review Press.

Trimi, S., & Berbegal-Mirabent, J. 2012. Business model innovation in entrepreneurship. *International Entrepreneurship and Management Journal*, 6: 449-495.

About the Authors

Patrick Brecht, M.Sc., works as a research associate at the Karlsruhe University of Applied Sciences after completing his study in industrial engineering with a focus on sales. In addition to his passion for sales, he has a growing fascination for science and entrepreneurship. In his role as project manager at the xLab, Experimental Lab for Innovation and Entrepreneurship, he supports startups in the area of business model innovation. His focus lies in validating startup potential within the field of smart business experiments and platform business models.

Daniel Hendriks, B.Sc., studied business administration and industrial engineering, focusing on service-oriented business model development at the Karlsruhe University of Applied Sciences. He works as a research associate at xLab of the University of Applied Sciences as an expert on B2B startup experimentation and Lean Startup. Besides his work, he is perusing a master's degree at the Karlsruhe Institute of Technology (KIT) in business administration and industrial engineering.

Anja Ströbele, B.Sc., is a research associate of the xLab at the Karlsruhe University of Applied Sciences. Her main research interests are in the fields of innovation, platform business models, and experimentation. During her studies, she developed a passion for research and an interest in innovation and entrepreneurship from analyzing the factors that turn a user innovator into a social entrepreneur with the help of theories of planned behavior. Her current master's degree focus is on marketing-finance that not only provides an understanding of the interface of these two disciplines, but also gives insights into design process models to create financial innovations. She currently works analyzing service management failures by applying probabilistic Bayesian modeling on collections of text documents.

Prof. Dr. Carsten Hahn works as a director of research and innovation at SAP and additionally holds a professorship for Innovation and Entrepreneurship at the University of Applied Sciences Karlsruhe. After studying business informatics at the University of Mannheim and doing his doctorate in Marketing at the University of Mainz, he began his career as an assistant to the executive board of SAP. Carsten serves as a visiting lecturer at the Sloan School of Management at the Massachusetts Institute of Technology. In his academic work, he founded the xLab, which deals with entrepreneurship and innovation concepts in research, teaching, and practical application.

Ingmar Wolff, M.Sc., is always passionate about innovation and entrepreneurship, has founded two startups, and financed his studies with a sales business. He completed his studies in industrial engineering with a focus on entrepreneurship and computer science at the Karlsruhe Institute of Technology (KIT). Afterwards, he consulted for startups on product-market fit and technology at the SAP startup incubator "innoWerft". After working at KIT as a research associate in the domain of automated machine learning, he founded his current company heliopas.ai GmbH, which provides AI-powered irrigation advice to farmers.

Citation: Brecht, P., Hendriks, D., Stroebele, A., Hahn, C.H., & Wolff, I. 2021. Discovery and Validation of Business Models: How B2B Startups can use Business Experiments. Technology Innovation Management Review, 11(3): 17-31.

http://doi.org/10.22215/timreview/1426

Keywords: Business experiments, growth hacking, business model, B2B startup, business experiment design, four-step iterative cycle, lean startup, customer development process

To address more complex social challenges, design thinking must become radically more collaborative and oriented toward systems change.

The Next Chapter in Design for Social Innovation

BY JOCELYN WYATT, TIM BROWN & SHAUNA CAREY

In 2013, Marie Stopes International (MSI) and the William and Flora Hewlett Foundation approached us with a challenge. As a global leader in reproductive health, MSI had spent decades delivering family-planning services to women in more than 30 countries around the world. But in several countries, including Zambia, the organization had struggled to reach unmarried, adolescent girls—a critical demographic in a country where many young women drop out of school because of unplanned pregnancy.

MSI sought to address this problem but wasn't sure where to start. And it wasn't alone—many global health organizations were also struggling to reach adolescents through tried-and-true programs. This type of challenge is ideally suited for human-centered design, a creative approach to problem-solving that starts with understanding people's needs and aspirations through immersive, participatory research and employs cycles of ideation, rapid testing, and refinement to develop solutions tailored to meet their needs. At IDEO.org, we apply human-centered design to a wide range of social challenges, from health equity to women's economic empowerment, to help nonprofits, community organizations, foundations, and governments design new, more effective ways to meet the needs of the communities they serve.

Over the course of several months, we partnered closely with the team at MSI Zambia to understand the barriers young women face in seeking and accessing reproductive health care, and to develop new

Photo courtesy of IDEO.org

During a live prototyping trip, a health worker leads a Smart Start counseling session with girls in rural Ethiopia.

models of delivery that could expand their access to information and contraceptive options. We began by speaking directly with young women and their influencers (parents, teachers, peers) and then reframing the challenge from their perspective. Through weeks of conversations and workshops, young women shared their mindsets, behaviors, and barriers related to dating, sex, and contraception. We contrasted what we heard from those young women with the perspectives of community members and with what we observed inside existing health centers. The gap between these perspectives began to reveal opportunities for design.

We quickly learned that if we wanted to reach more teens, we needed to design an environment for conversations about sexual health that felt less risky and more relatable to young women. They found sterile clinics with waiting rooms full of older, married women and messages about "family planning" uninviting; they were curious about sexual health but not yet planning a family. This insight led to a series of prototypes that we tested in the community over a number of weeks. One early prototype—a pop-up health session at a local nightclub—failed miserably. Teens weren't interested in having personal conversations interrupt their night out. Another prototype—a pop-up nail salon staffed by MSI's nurses—was much more effective. When a manicure was a focal point helping to diffuse the intensity of personal conversations, girls were more willing to open up about their biggest questions concerning love and sex, and they were able to pause for breaks to select polish colors. We saw that instead of infusing relatable elements into the clinic experience, we needed to reimagine it altogether.

This early prototype eventually evolved into the Diva Centres—a set of teen-friendly spaces where girls can hang out with their friends and do their nails while having informal conversations about dating, relationships, and sex. If you entered one of these centers, you'd likely find a group of young women chatting casually over small tables set with magazines and brightly colored polish—nearly unrecognizable as a clinic, except for the friendly nurse available to answer questions if and when girls are ready. And it's not just the physical environment that creates a safe and welcoming experience. Peer educators serve as the program's ambassadors out in the community. And providers have been trained to deliver teen-friendly messaging and a judgement-free experience for their young clients, using materials that feature aspirational female characters we designed—the Divas—to demystify and personify choices about birth control for young women, sparking curiosity instead of fear.

By many measures, these Diva Centres succeeded. They spurred a 16-fold increase in adolescents served by MSI Zambia and a stunning 84 percent uptake in contraception among young women after they engaged with the service. These early outcomes helped demonstrate the power of design in creating more engaging, relevant services and played an important role in shifting conversations about adolescent sexual and reproductive health to center more on teen girls' unique needs and aspirations.

JOCELYN WYATT is the cofounder and CEO of IDEO.org, a nonprofit design studio whose mission is to create a more just and inclusive world.

TIM BROWN is chair of global design and innovation firm IDEO and vice chair of kyu.

SHAUNA CAREY is a managing director at IDEO.org.

But the project, in retrospect, also provides an interesting benchmark for the many ways in which the application of human-centered design to social problems has evolved in recent years. At IDEO.org, we've had a front-row seat to this transformation over the past decade. We've engaged in more intimate, hands-on collaboration with communities to surface better solutions. We've embedded more experimentation, iteration, and adaptation into the development of new programs and services. We've seen how embracing a more rigorous approach to planning for scale and sustainability has produced superior and more enduring outcomes. In short, the practice of human-centered design has advanced to become better able to address the systemic change that the social sector seeks.

FROM NASCENT PRACTICE TO RICH ECOSYSTEM

In 2010, one year before IDEO.org was founded, we wrote a cover story for this magazine about the critical role design could play in driving positive impact in the social sector.[1] Tim Brown headed IDEO, the global design consultancy, while Jocelyn Wyatt led its new social innovation group, which would become IDEO.org. Businesses in the private sector were widely adopting design thinking as a means of driving innovation, but nonprofits and social enterprises had only just begun to incorporate the methods and mindsets of design—mostly in one-off projects. It was a pivotal moment for the social sector, still reeling from the 2008 recession.

The design-led approach to social innovation has gone from a nascent practice to a rich ecosystem of adopters around the world.

We believed in the potential of design to create more resonant, relevant, and integrated solutions to a wide range of challenges related to poverty and inequality. So we made the case to social-sector organizations that a design-led process could help them meet the needs of those they serve. The social sector already embraces an ethos of focusing on end users, but human-centered design, we argued, could unlock new possibilities for creative solutions to social problems.

Although there are many ways to describe human-centered design, perhaps the simplest is to focus on three, sometimes overlapping, phases of problem-solving: *inspiration*, *ideation*, and *implementation*. These phases build on each other, although they are not necessarily linear. In the inspiration stage, designers engage in

PHOTO COURTESY OF IDEO.ORG

Peer Connectors play a crucial role in the outreach strategy of the Diva Centres by helping girls learn about contraception in their own terms.

immersive, participatory research with the ultimate users of a new product or service (whether a business or communities affected by the problem at hand) in order to surface insights about their needs, barriers, and aspirations. During ideation, we begin to develop concepts that respond to the needs uncovered during research and test them through various rounds of prototyping. Prototypes are tangible experiments conducted quickly, inexpensively, and on a small scale. We design them to help validate (or invalidate) potential solutions, as well as to uncover unforeseen challenges with implementation. Finally, during the implementation phase, we refine solutions that showed promise during testing as we build the delivery model, technology, tools, and processes needed to ensure long-term success.

Over the past decade, this design-led approach to social innovation has gone from a relatively nascent practice championed by a few organizations to a rich ecosystem of practitioners, academics, and adopters around the world. At IDEO.org, we've had the opportunity to use design to address a vast array of challenges—from working with mobile-network operators in South Asia and East Africa to create digital financial services that advance women's financial inclusion to partnering with institutions to address racial inequities in health care.

Design organizations around the globe are now working on a range of social challenges. The MASS Design Group is establishing first-rate hospitals and a medical campus in one of the poorest districts of Rwanda, D-Rev is developing affordable prosthetics for low-resource environments, and the Vihara Innovation Network is operating a human-centered design practice in rural Bihar, India. And communities of practice that bring together practitioners, funders, and designers—such as the global health collective Design for Health and HCD Exchange, which focuses on adolescent sexual reproductive health in sub-Saharan Africa and South Asia—have emerged to connect practitioners, funders, and organizations looking to apply design in their worlds.

This surge in interest in the social sector also inspired us, in 2009, to create Design Kit (originally published as HCD Toolkit), a suite of tools and courses—hosted with Acumen Academy—to help social-impact organizations adopt design methods in their own work. Since then, we've seen more than one million users access these tools from countries around the world.

But as the practice of design has become more mainstream, the approach has also generated a fair amount of criticism. Objections have focused chiefly on the up-front costs of design and its potential to produce unscalable, bespoke solutions without sufficient evidence underpinning the intervention. Early cases in our own portfolio of work, including the Diva Centres, fed into these concerns.

While the initial sites of the Diva Centres showed results in reaching adolescent girls and doubling the proportion of adolescent clients selecting a long-acting contraceptive method, we saw that the model would struggle to expand beyond the pilot. Our design process had focused too narrowly on creating an engaging and effective user experience for teens and providers, without grounding it in the wider health ecosystem. Because the stand-alone centers weren't integrated with existing service-delivery channels, they were prohibitively expensive and complicated to scale.

In the years since the Diva Centres were piloted, MSI Zambia has worked with Spring Impact, a consultancy that helps organizations scale social-impact solutions, to integrate the model into public-health facilities. The partnership has adopted lessons from the pilot program and developed a more iterative and embedded program model that is still in use today. But we could have avoided some of these adaptation efforts by supplementing our design process. A more intensive, up-front mapping of the surrounding health-care ecosystem and a deeper assessment of our implementing partners' capacity would have enabled us to design a more cost-effective and scalable intervention from the start.

In a subsequent design project with MSI Burkina Faso in 2017, our team put these lessons into practice. We worked closely with MSI's local staff to understand the service delivery ecosystem and the organization's capacity constraints from the start. The resulting program we designed together, La Famille Idéale, augmented and built upon that foundation. This more rigorous approach to the design process enabled us to rapidly adapt the program for implementation in other countries across the Sahel region. These practices are a standard part of our design process today but were

not firmly cemented when we began the Diva Centre project a few years earlier.

EVOLVING THE PRACTICE OF DESIGN

While cases like the Diva Centres helped popularize human-centered design in the social sector, they also highlighted opportunities for the practice to progress in addressing complex social problems. Challenges in health, education, migration, and economic inclusion are often described as "wicked problems," because they are difficult to define, involve multiple interconnected actors, and likely require more than one intervention. While design can be an effective tool for developing novel and successful interventions, many design efforts are not effectively structured to tackle these problems, because inflexible funding, time scales, and ways of working inherited from the private sector limit them.[2]

In recent years, many human-centered designers have structured initiatives in new ways to meet these complex social challenges. One example of this approach is the Adolescents 360 (A360) program, which also focused on supporting more young women to make decisions about their reproductive health but, thanks to lessons learned from Divas, took a more multidisciplinary, adaptive, and ultimately scalable approach.

A360, led by global health nonprofit PSI and funded by the Bill & Melinda Gates Foundation and the Children's Investment Fund Foundation (CIFF), seeks to reframe how girls ages 15 to 19 perceive, value, and access modern contraception across Tanzania, Ethiopia, and Nigeria. The project, which launched in 2016, has several advantages over previous design-led initiatives for adolescent health. It was formed by a transdisciplinary network of partners that integrates human-centered design with expertise from public health, adolescent developmental science, cultural anthropology, and social marketing. This combined approach aims to produce youth-friendly solutions that are poised to scale. IDEO.org was the human-centered design partner within the consortium.

The A360 program would need to build off a rich market analysis and evidence-based theory of change to have a shot at meeting its goals. To ground the design process, A360 conducted a rigorous review of relevant academic literature and evidence before beginning user research. It sought to serve a whopping 60,000 to 70,000 girls in rural Ethiopia alone within three years. The market analysis helped us target our qualitative design research and ideation with a greater awareness of the contextual factors affecting girls' lives and choices.

A360 also recruited, trained, and worked alongside on-the-ground experts: Young women and men from local communities in Ethiopia joined our team as researchers and co-designers throughout the process.

And the network of partners wasn't the only thing that differentiated A360's design process from its predecessors'. The consortium set up the program as a multiphase, 15-month project—recognizing that a longer process was required than had been available for the Diva Centres—that allowed for multiple cycles of exploration, prototyping, refinement, measurement, and optimization, hand in hand with PSI, to ensure that our solution not only succeeded in the pilot phase but also had a clear pathway to scale.

As the qualitative design research for A360 in Ethiopia began, the process didn't look substantially different from that of the Diva Centres or other previous projects. Ethnographic research provided insights about young people's behavior that signaled opportunities for innovative design. It quickly became clear that strict social norms tightly govern girls' movement and access to services in these extremely rural communities. Any viable solution would need to build support for contraception among young women's immediate network of influencers, including elders in the community and—in most cases—husbands. We found from speaking to young men and women that they were more open to new ideas and possibilities for the future immediately after marriage—when the prospect of planning their lives together was most relevant.

With this insight in hand, the team prototyped different ways of broaching the topic of contraception that seized this moment as an entry point for expanding girls' access to services. The team tested several concepts, including a financial planning pop-up at the local market, a marriage counseling session, and a cost calculator to support couples in determining the resources they would need to plan for the families they wanted. The cost calculator resonated most among test users, who expressed anxiety about their financial future. This observation (among others) gradually reshaped the framing of the proposed service—from a conversation about family planning (taboo in the local context) to one about financial planning, in which contraception and birth spacing became tools introduced later to support young couples in achieving their goals for financial security. These prototypes were gradually refined into a cohesive, holistic program called Smart Start that is being implemented across Ethiopia today.

Smart Start frames contraception as the leeway necessary to save up for a home before a firstborn, as the difference between comfortably raising one child and making ends meet with two, and as a decision couples can navigate together. To enable young men and women to participate effectively in these sessions, despite low literacy, the design team explored many ways of communicating these financial trade-offs in simple, accessible terms. Eventually, this search led to the discovery that bags of teff—the country's staple crop—could act as a universally understood proxy for money.

Smart Start sends health-care workers to meet couples where they are. The delivery model leverages and augments the country's network of health extension workers (HEWs), which routinely sends trained health-care providers to rural communities to deliver frontline health-care services. Through Smart Start, these HEWs are trained to lead counseling sessions with couples that begin with a conversation about life goals designed to help young married girls and their partners consider how their long-term visions will be more attainable if they delay their first pregnancy and space out subsequent births. HEWs are equipped with engaging visual materials and interactive exercises that help couples think through the costs of raising a child, and present contraception as a tool that can help them align their financial and reproductive plans. And before counseling begins, community meetings are held, where local leaders and elders discuss the importance of setting up young couples with the tools they need to build a better future. This step gives these couples sanction to explore an otherwise taboo conversation.

A360 also partnered designers with an adolescent developmental scientist to develop communication tools tailored to young people's stages of cognitive and hormonal development. Health counselors reported that the thoughtful messaging helped increase understand-

ing and connection with young clients who were previously difficult to reach. After one conversation with a Smart Start counselor, girls can easily memorize the information and then share it with parents and other community gatekeepers, helping them to understand the logic behind choosing to delay pregnancy.

Since Smart Start's pilot, more than 30,000 adolescent girls have adopted modern contraception through the program. PSI's results exceeded expectations and convinced Ethiopian health ministers to visit the pilot and learn more. The A360 model was created with built-in systems that would produce evidence that the Ministry of Health deemed compelling. These systems included careful program monitoring, led by PSI country teams, and independent process and outcome evaluations led by measurement partners Itad and the London School of Hygiene & Tropical Medicine. The conversion was remarkable. Based on these results, the ministry has committed to expand the Smart Start program across the country, with $10.5 million from the Children's Investment Fund Foundation. The program will be embedded in the country's primary health-care system and will support more than one million girls by the end of 2024.

Human-centered design has a role to play in shifting our communities toward a more equitable and regenerative future.

The adoption and integration of Smart Start into the national government health-care system is one of several critical metrics of success: It demonstrates the potential for scalability, financial feasibility, and long-term operational sustainability. Most important, it also shows that the program works for the young people it was designed for.

THREE ACTIONS FOR BETTER DESIGN

The lessons learned from Smart Start and other similarly ambitious design programs over the past several years have helped inform an era of social-impact design that is increasingly rigorous, experimental, and collaborative. The field has made tremendous progress. But as we look toward the next 10 years, our practice must continue to evolve in several essential ways.

Invest in the infrastructure and evidence required to scale. | While many social-impact design efforts over the past decade have demonstrated innovation in tackling social challenges, far fewer have demonstrated change at scale. Like the Diva Centres, early case studies of design for social innovation focused on creating demonstration cases and more engaging user experiences, but often lacked the necessary ingredients to create large-scale change.

In many instances, this deficiency springs from a failure to design the "back of house" organizational infrastructure required to bring bold new products or services to life. Across the board, leading social-impact designers—including us for many years—failed to recognize and solve for the significant operational and mindset shifts required for an organization to deliver effectively a new and innovative intervention. This blind spot leads to solutions that are vulnerable to challenges of an organization's operational capacity, capabilities, and funding. Preparing an implementation plan, business model, role descriptions, and measurement frameworks can help overcome these barriers and steward a solution to success, but they are not widely integrated into the practice of social-impact design across the sector.

Consider, for example, Kuja Kuja, a real-time customer feedback platform that enables refugees to rate their satisfaction with the services NGOs deliver. The platform idea belongs to Alight (formerly the American Refugee Committee), which sought to reimagine the ways in which humanitarian organizations improve and take responsibility for their services and to shift the role of refugees from beneficiaries to customers. Today, staff at Alight HQ can see photos of service points, read individual testimonials, and gather qualitative information before making decisions. Kuja Kuja has collected more than one million data points in six countries.

But getting the service off the ground was only a small part of what we had to design and deliver. Creating the tools and conditions to foster organizational change presented the major design challenge. We had to answer questions such as: How can we build a trustworthy brand, given the power dynamics at play between refugees and NGOs? How can the feedback from the Kuja Kuja platform improve daily operations at the camp? How will this system work when network connectivity goes out? How will we recruit and train local teams to do this work? Developing elegant training materials was as important as getting the digital interactions right, and laid the groundwork for Kuja Kuja to scale quickly. It expanded to five countries two years after piloting in a single refugee settlement in Uganda.

Getting a solution to market is only the first step. Designers must also plan for phases of intentional learning, optimization, and refinement beyond the pilot and build measurement systems into their design solutions to produce the evidence needed for scaling. The Smart Start program in Ethiopia illustrates this approach. It has scaled deliberately and incrementally from live prototype to pilots to regional expansion. This journey included additional phases of design as the program expanded to iterate, optimize, and improve the cost-effectiveness of the model at each phase of growth. For example, one year after the initial pilot, PSI and IDEO.org used a six-week design sprint to reduce costs and lighten the workload for HEWs that implement the model, so that the Ministry of Health, which plans to integrate it into the national health system, could adopt it more easily. Success will depend on ensuring not only that the experience offered to young couples continues to be positive, but also that any adaptations to the delivery model feel meaningful and efficient for the HEWs who need to execute it.[3] Formalizing optimization as a distinct phase of design can help implementing organizations move

more confidently from pilot to replications and scale.

This patient, phased approach to scaling innovation also requires new approaches to funding. Scaling a well-designed and proven intervention still requires a significant investment—usually from the public sector or philanthropy. In the past several years, we've seen the rise of systems-change philanthropy, which aims to join the forces of individuals, organizations, and institutions to address the root causes of social problems. New, ambitious investment vehicles designed to support such philanthropy—through initiatives such as Co-Impact, the MacArthur Foundation's 100&Change program, the Global Innovation Fund, and The Audacious Project—have arisen to satisfy this demand. But many design-led interventions have yet to successfully capitalize on this funding because of the evidence gap many well-designed solutions face. They have yet to scale and have not been properly evaluated by third parties.

A group of women huddle around a family planning board game in Niger during the adaptation design for La Famille Idéale.

As human-centered design has gained popularity as a tool for developing new delivery models, the social sector has invested significantly in evaluations to collect evidence about the methodology itself, but much less in outcome evaluations for design-led solutions. In order to build the case to replicate and scale such solutions, design firms need to collect the evidence necessary to demonstrate whether the solutions have generated the desired outcomes. Firms need to partner with experts and embed detailed frameworks for measurement and learning in all services and products they design as they are implemented. Without this commitment, many transformative, human-centered solutions will be unnecessarily limited in reach.

Build community ownership. | We must also invest in the people we serve, by yielding power to them in the design process. Human-centered design has always been a collaborative process, in which designers, subject-matter experts, and communities themselves are all engaged in finding and developing solutions. But as design continues to take root in the social sector, these exchanges must shift from mere consultation to co-creation, in which those who will ultimately benefit from and/or deliver a solution can shape the outcome at each stage of the design process. When the stakes of design change (to improve people's health, livelihood, or well-being), our process must follow suit—ensuring that communities have a voice and a hand in bringing possibilities to life.

This approach requires us to design the conditions for co-creation. We must build a space that enables those with lived experience of the issue at hand to participate fully. "Good design makes space for those without formal training to shape and control the project itself," says George Aye, of the social-impact design firm Greater Good Studio. "We must design the conditions where the constituents own the change we're asking them to make."[4]

We had good intentions for the Diva Centres. We conducted participatory research with teen girls in the program's target population. We created daily diaries and other tools to help them articulate their needs, goals, and barriers. We also tested concepts with them to see what resonated. But as we refined those concepts toward a final solution, the closeness of our interaction faded, and the young women who helped inspire the Diva Centres had a diminished role during the later phases of the process. As a result, some of the final program materials contained depictions that didn't reflect the girls they were created for and needed to be modified after implementation had begun.

Later on, when we worked with MSI to design an adaptation of the Divas program for a rural audience, MSI's field team in Zambia took the lead. Building on the lessons learned from that initial phase of work, MSI staff led design activities with girls and their families directly, working with our designers to visualize potential directions for the brand and materials and ensuring that communities remained engaged throughout the process.

In the case of Smart Start, the program was intentionally co-designed from the beginning with young people and local PSI staff, who had personal experience of the challenge at hand. Together, we did several rounds of refinement on the program's tools and communications materials. We ended up with visual representations of people and community that resonated, and used bags of grain as a proxy for money that everyone in the village could relate to their own household finances. Everything in the program's outreach materials, from the color patterns used to the clothes worn in the illustrations, expresses the hyperlocal context.

PHOTO COURTESY OF IDEO.ORG

The success of Smart Start belies an important truth: The best design happens with communities, not *for* them. Consequently, we must always focus on creating the conditions for the people affected most directly by an issue to articulate the solutions they want and need, and use our skills as designers to help bring those initiatives to life. Organizations such as Creative Reaction Lab in St. Louis, the National Equity Project, and the K12 Lab at the Hasso Plattner Institute of Design at Stanford University (the d.school) are crafting new tools[5] and publishing resources (such as the *Field Guide to Equity-Centered Community Design* and the Liberatory Design Resource Collection) to help others practice more inclusive and equitable design processes. For example, they are training young people in underinvested communities to lead the design process and codifying practices and frameworks that recognize the role of power and historical context in the design process.

In recent years, new educational programs have also been established to support a more diverse cadre of young people in finding career pathways in design. Typically, design education has produced a field of practitioners that are disproportionately white, male, and privileged. By contrast, most social-sector design—by virtue of its orientation toward social justice—focuses on challenges that disproportionately affect women, people of color, and other historically marginalized groups. Today, programs such as Nairobi Design Institute in Kenya and Inneract Project in the United States are helping to narrow that gap, by providing training and support to designers with a much wider set of lived experiences to draw upon in imagining and creating new solutions. This more inclusive vision for design, with a plurality of perspectives at the table, will undoubtedly yield better results—especially within the social sector.

Embrace radical collaboration focused on systemic change. | The major social problems we face today are grounded in systemic issues that demand serious reengineering—from staggering income inequality to waste and exploitation in our global food system to inequitable and overstretched health-care systems around the world. Human-centered design has a role to play in shifting our communities toward a more equitable and regenerative future, but it will look quite different from the social innovation that has come before. Instead of focusing on designing individual products and services in response to discrete briefs, designers must adopt a more systemic lens and intentionally coordinate multiple interventions with multiple actors to create more enduring change.

We've been on this journey with our own portfolio of reproductive-health work. In global development, we know that when young people can control whether and when to have children, rates of unintended pregnancies and maternal mortality decrease dramatically. Delaying pregnancy can also enable girls to finish secondary school and pursue further education and work opportunities before starting a family—and thereby gives them more agency in shaping the future they want. Since 2014, IDEO.org has worked with partners in more than 10 countries to expand young women's access to relevant, accessible reproductive-health care. Together, the resulting programs have led to more than 400,000 young women adopting a contraceptive method of their choice.

But from a girl's perspective, the relevant issues are not so simple. Barriers in other parts of girls' and women's lives, such as lack of opportunities to earn income and a limited sense of personal agency and empowerment, often prevent them from benefiting fully from the promise of contraception. While the social sector is frequently divided into actors that address these issues individually, girls themselves experience them holistically. For a teenage girl, access to contraception and choices about sex and relationships are intrinsically linked to the other areas of her life. Creating systemic change in the opportunities young women have to shape their lives will require coordinated efforts within and beyond access to reproductive-health care. This, in turn, demands that many types of actors, from NGOs and foundations to private-sector companies, community activists, and adolescent girls themselves, work together.

Today's social sector is not set up for this type of collective problem-solving; its incentive structures demand narrow outcomes at scale and revolve around competition for funding. To transcend these limitations, we must find new ways to enable radical collaboration and new vehicles for change.

One approach we're trying is the Billion Girls CoLab. The CoLab brings together cross-sectoral cohorts of organizations to work together on a range of issues that affect young women's health, well-being, and economic opportunities. Together with adolescent girls as partners, these organizations work to identify shared problems and test new solutions through collaborative design sprints. We kicked off the CoLab in January 2020 with a "make-a-thon," or collaborative design workshop, in Nairobi, Kenya. It produced 10 wide-ranging prototypes for new girl-centered services that address a wide range of issues, from body positivity and menstrual health to helping girls find employment opportunities.

The Billion Girls CoLab is still in its early days, but its promise is exciting. Its goal is to take the field of systems change, which has often been the domain of academia, and make it actionable by creating spaces that break sector silos, promote more holistic designs, and transfer power into the hands of users—in this case, adolescent girls. This type of design requires a significant shift in the role of designers, from solution creators to system architects; bringing together the right mix of organizations and stakeholders; and creating the conditions and tools to facilitate productive, inclusive collaboration. This shift also invites private- and public-sector actors to pool learnings, share resources, and orchestrate change in a more coordinated and collaborative manner. A multisector initiative of this magnitude, although complex, can unlock much more significant, systemic change, where the creativity and ingenuity of the design process is most needed.

"Systems change all the time," says Dave Kim, program officer on financial services for the poor at the Bill & Melinda Gates Foundation. "It is only when we apply a design lens—one that leaps across disciplines while remaining grounded in human insights—that we can tilt the change to a system that is a bit more just and joyful." ■

Notes

1 Tim Brown and Jocelyn Wyatt, "Design Thinking for Social Innovation," *Stanford Social Innovation Review*, vol. 8, no. 1, 2010.

2 Chris Larkin, "Design for Impact: Bringing Empathy and Creativity to Social Change," in Nina Montgomery, ed., *Perspectives on Impact: Leading Voices on Making Systemic Change in the Twenty-First Century*, London: Routledge, 2019.

3 Ibid.

4 George Aye, "It's Time to Define What 'Good' Means in Our Industry," *Design Observer*, March 14, 2019.

5 Creative Reaction Lab, *Field Guide: Equity-Centered Community Design*.

Accra, Ghana, August 2026