



CLAWAR Association Series on  
Robot Ethics and Standards

# Sustainable Ethics and Values in the Design and Regulation of Robotics and AI

## Editors

Roeland W. de Bruin  
Mohammad O. Tokhi  
Lucky Belder  
Maria Isabel A. Ferreira  
Naveen S. Govindarajulu  
Manuel F. Silva



# **SUSTAINABLE ETHICS AND VALUES IN THE DESIGN AND REGULATION OF ROBOTICS AND AI**





# **SUSTAINABLE ETHICS AND VALUES IN THE DESIGN AND REGULATION OF ROBOTICS AND AI**

**ICRES 2023 Proceedings,  
Utrecht, The Netherlands, 17-18 July 2023**

Editors

**Roeland Wieger de Bruin**

*Utrecht University, Netherlands*

**Mohammad Osman Tokhi**

*London South Bank University, UK*

**Lucky Belder**

*Utrecht University, Netherlands*

**Maria Isabel Aldinhas Ferreira**

*University of Lisbon, Portugal*

**Naveen Sundar Govindarajulu**

*Rensselaer Polytechnic Institute, NY, USA*

**Manuel F. Silva**

*Porto Polytechnic, Portugal*

*Published by*

CLAWAR Association Ltd, UK ([www.clawar.org](http://www.clawar.org))

Sustainable ethics and values in the design and regulation of robotics and AI  
Proceedings of the Eighth International Conference on Robot Ethics and  
Standards

Copyright © 2023 by CLAWAR Association Ltd

ISBN 978-1-7396142-1-8

## **PREFACE**

ICRES 2023 is the eighth edition of the annual International Conference series on Robot Ethics and Standards. The conference is organized by CLAWAR Association in collaboration with the Utrecht University, Focus Area on Governing the Digital Society, Special Interest Group Principles by Design: towards good data practice, and the RENFORCE research group, School of Law in Utrecht, The Netherlands during 17 – 18 July 2023.

ICRES 2023 brings new developments and new research findings in robot ethics and ethical issues of robotic and associated technologies. The topics covered include fundamentals and principles of robot ethics, social impact of robots, human factors, regulatory and safety issues of robotics and artificial intelligence.

The ICRES 2023 conference includes a total of 20 regular and special session presentations, and seven invited and plenary lectures delivered by worldwide scholars. This number has been arrived at through rigorous peer review process of initial submissions, where each paper initially submitted has received on average three reviews. The conference additionally features special sessions on AI regulations, human-robot collaboration in work environments, and a thematic student-led workshop.

The editors would like to thank members of the International Scientific Committee and National Organising Committee for their efforts in reviewing the submitted articles, and the authors for addressing the comments and suggestions of the reviewers in their final submissions. It is believed that the ICRES 2023 proceedings will be a valuable source of reference for research and development in the rapidly growing area of robotics and associated technologies.

R. W. de Bruin, M. O. Tokhi, L. Belder, M. I. A. Ferreira, N. S. Govindarajulu, and M. F. Silva,

## CONFERENCE ORGANISERS



**CLAWAR Association**  
[www.clawar.org](http://www.clawar.org)



**Utrecht  
University**

**Utrecht University**  
<https://www.uu.nl/>

## CONFERENCE SPONSORS AND SUPPORTERS



<https://www.kienhuishoving.com/>

## CONFERENCE COMMITTEES AND CHAIRS

### Conference Chairs and Managers

Roeland Wieger de Bruin (General Co-Chair)	– Utrecht University, Netherlands
Mohammad Osman Tokhi (General Co-Chair)	– London South Bank University, UK
Maria Isabel Aldinhas Ferreira (General Co-Chair)	– University of Lisbon, Portugal
Gurvinder S. Virk (Co-Chair IAC)	– CLAWAR Association, UK
Endre E. Kadar (Co-Chair IAC)	– University of Portsmouth, UK
Naveen S. Govindarajulu (Co-Chair ISC)	– Rensselaer Polytechnic Institute, USA
Lucky Belder, (Co-Chair ISC)	– Utrecht University, Netherlands
Manuel F. Silva (Co-Chair ISC)	– ISEP-IPP and INESC TEC, Portugal
Roeland W. de Bruin (Organising Committee Co-Chair)	– Utrecht University, Netherlands
Endre E. Kadar (Organising Committee Co-Chair)	– University of Portsmouth, UK

### International Advisory Committee

Endre E. Kadar	– University of Portsmouth, UK
Gurvinder S. Virk	– CLAWAR Association, UK
Selmer Bringsjord	– Rensselaer Polytechnic Institute, USA
Raja Chatila	– ISIR/UPMC-CNRS, France
Jen-Chieh Wang	– Industrial Technology Research Institute, Taiwan
Alan Winfield	– University of West England, UK

### Organising Committee

Roeland W. de Bruin	– Utrecht University, Netherlands
Endre E. Kadar	– University of Portsmouth, UK
Abdullah Almeshal	– College of Technological Studies, Kuwait
Lucky Belder	– Utrecht University, Netherlands
Maria Isabel Aldinhas Ferreira	– University of Lisbon, Portugal
Sarah Fletcher	– Cranfield University, UK
Gabriela Gallegos Garrido	– London South Bank University, UK
Khaled Goher	– University of Nottingham, UK
Naveen S. Govindarajulu	– Rensselaer Polytechnic Institute, USA
Philip Lance	– PA Consulting, UK
Manuel Silva	– ISEP-IPP and INESC TEC, Portugal
Mohammad Osman Tokhi	– London South Bank University, UK
Louie Webb	– London South Bank University, UK
Karolina Zawieska	– Industrial Research Institute for Automation and Measurement, Poland

### International Scientific Committee

Lucky Belder	– Utrecht University, Netherlands
Naveen S. Govindarajulu	– Rensselaer Polytechnic Institute, USA
Manuel Silva	– ISEP-IPP and INESC TEC, Portugal
Ronald Arkin	– Georgia Institute of Technology, USA
Fabio Bonsignorio	– University of Zagreb, Heron Robots, Italy

Selmer Bringsjord	– Rensselaer Polytechnic Institute, USA
Sunyong Byun	– Seoul National University of Education, South Korea
Raja Chatila	– ISIR/UPMC-CNRS, France
Sarah Fletcher	– Cranfield University, UK
Gabriela Gallegos Garrido	– London South Bank University, UK
Moojan Ghafurian	– University of Waterloo, Canada
Khaled Goher	– University of Nottingham, UK
Yue Hu	– University of Waterloo, Canada
Shin Kim	– Hankuk University of Foreign Studies, South Korea
Philip Lance	– PA Consulting, UK
Rosa Lee	– Korean Information Society Development Institute, South Korea
Tim Cheongho Lee	– Sangmyung University, South Korea
Fiorella Operto	– Scuola di Robótica, Italy
Amit Kumar Pandey	– SoftBank Robotics, France
Edson Prestes	– Universidade Federal do Rio Grande do Sul, Brazil
Matthew Studley	– University of the West of England, UK
David Vernon	– Carnegie Mellon University Africa, Rwanda
Gianmarco Veruggio	– National Research Council, Italy
Jen-Chieh Wang	– Industrial Technology Research Institute, Taiwan
Alan Winfield	– University of West England, UK
Sule Yildirim-Yayilgan	– Norwegian University of Science and Technology, Norway
Karolina Zawieska	– Industrial Research Institute for Automation and Measurement, Poland





## TABLE OF CONTENTS

Title .....	i
Preface .....	vii
Conference organisers .....	viii
Conference sponsors and supporters .....	ix
Conference committees and chairs .....	x
Table of contents .....	xiii

### Section–1: Plenary presentations

Governing generative AI: A disinformation perspective .....	3
<i>Madeleine de Cock Buning</i>	
AI and human rights – the need for global standards .....	4
<i>Jan Kleijssen</i>	
How hard, logico-mathematically speaking, is real war (including of the ethically correct variety) for an AI: Answers from study of the Game Ekte Krig .....	5
<i>Selmer Bringsjord, Naveen Sundar Govindarajulu and Alexander Bringsjord</i>	
How do you lip read a robot-mitigating the risks triggered by the procurement of AI powered HR technology .....	6
<i>Susan Scott-Parker</i>	
Benchmarking, robots, and smart cities .....	9
<i>Matthew Studley</i>	

### Section–2: Regular Papers

Embedding public values in an online public community network: A scenario-based approach .....	13
<i>Mathilde Sanders, Erna Ruijter and Josè Van Dijck</i>	
Introducing the concept of repurposing robots; to increase their useful life, reduce waste, and improve sustainability in the robotics industry .....	16
<i>Helen McGloin, Matthew Studley, Richard Mawle and Alan Winfield</i>	
Non-profit-driven development paths for equitable AI/AGI .....	28
<i>Michael Giancola and James Oswald</i>	
Roboethics in the loop: The ELS issues in the REXASI-PRO project .....	39
<i>Gianmarco Veruggio, Maurizio Mongelli and Fiorella Operto</i>	
Prospective insights from the rear-view mirror: Revisiting the MONARCH project .....	47
<i>Maria Isabel Aldinhas Ferreira</i>	
Unaware self-nudging by social robots .....	55
<i>Stefano Calboli</i>	

Sock it to challenging behaviour - Development of Smartsocks technology for early detection of distress in people with neurological disorders: a survey study .....	64
<i>Iveta Eimontaite, Zeke Steer, Jacqui Arnold and Patrick Ogundele</i>	

### **Section–3: Human-Robot Collaboration**

Towards an ethical framework for human-robot collaboration in manufacturing: Methodological considerations .....	79
<i>Tiziana C. Callari, Ella-Mae Hubbard, and Niels Lohse</i>	

How can human-robot collaboration improve operators’ working conditions and wellbeing in aircraft fuel tank inspection: A mixed-methods user-centred approach .....	90
<i>Vaishnavi Sashidharan, Iveta Eimontaite, Sarah R. Fletcher, Nikos Dimitropoulos, Sotiris Makris, George Michalos, Igal Israeli and Scott Tucker</i>	

To collaborate or not to collaborate? How to determine the most suitable level of automation to increase workforce sustainability and production efficiency .....	99
<i>Sacha Godhania, Iveta Eimontaite, Sarah R. Fletcher, Ana Gonzalez Segura, Eloy Ruiz and Ana Paola Caro Ospina</i>	

Developing unmanned aerial robotics to support wild berry harvesting in Finland: Human factors, standards and ethics .....	109
<i>Sarah R. Fletcher, Anne-Marie Oostveen, Paul Chippendale, Micael Couceiro, and Laura Smith Ballester</i>	

### **Section–4: AI Regulations**

AI, GLAM, Innovation and trust: Towards value-based regulation by design .....	123
<i>Kelly Breemen and Vicky Breemen</i>	

Proof of causation under uncertainty .....	124
<i>Elbert de Jong</i>	

The governance of generative AI: Observability, modifiability, accessibility .....	125
<i>Fabian Ferrari</i>	

Privacy aspects of AI regulation .....	126
<i>Lesley Broos</i>	

### **Section–5: Workshop: How do you want to be governed? Should we pace innovation? A ChatGPT Case Study**

How can we compare llms in their current state with the human capability of understanding? .....	129
<i>Luiza Świerzawska</i>	

Should we call the hallucinating oracle an expert?: Large language models and the concept of expertise .....	135
<i>Thomas Wachter</i>	

Making human progress or falling behind: Does ChatGPT exacerbate the knowledge gap? .....	141
<i>Jiayi Liu</i>	

The new AI era: An amazing fantasy(?) .....	147
<i>Panagiota Rassia</i>	

Is generative ai depriving humans of creativity? .....	152
<i>Alessia Giulimondi</i>	
Author index.....	157



**SECTION-1**  
**PLENARY PRESENTATIONS**



## **GOVERNING GENERATIVE AI: A DISINFORMATION PERSPECTIVE**

MADELEINE DE COCK BUNING

*School Law, Utrecht University, The Netherlands*

Generative AI, such as brought to society's attention by means of ChatGPT, is promising in many ways. Given its abilities, Generative AI could greatly contribute to the availability and accessibility of knowledge within all layers of society. At the same time, the outputs of Generative AI can potentially spread disinformation with impactful consequences for societies when it would be taken for "truth".

## **AI AND HUMAN RIGHTS – THE NEED FOR GLOBAL STANDARDS**

JAN KLEIJSEN

*Directorate General Human Rights and Rule of Law, European Commission*

This presentation will address the fast developments of AI, and the faith that society puts therein – as opposed to the imminent and significant risks AI might pose to human civilization. A scenario that comes to mind, is the one in which human creativity is entirely replaced by AI. The presentation (rhetorically) questions the impact on individuals and society as a whole – and the functioning of the democratic system as we know it. Whereas AI can be promising in many ways, regulating the potential downsides of uncontrolled application of AI is urgent. Fortunately, great efforts are already being made – including those by the OECD, the European Parliament and Unesco to name a few. Illustrating the ongoing regulatory work in general, Jan will specifically focus on the work that is currently being done by the Council of Europe.



## **HOW HARD, LOGICO-MATHEMATICALLY SPEAKING, IS REAL WAR (INCLUDING OF THE ETHICALLY CORRECT VARIETY) FOR AN AI: ANSWERS FROM STUDY OF THE GAME Ekte KRIG**

SELMER BRINGSJORD, NAVEEN SUNDAR GOVINDARAJULU AND ALEXANDER  
BRINGSJORD

*Rensselaer Polytechnic Institute, New York, USA*

Point #1: Paul Scharre (2023) is correct that AI is the pivotal military battleground of the 21st century. Point #2: The field of AI has long focused on games; Checkers, Chess, and Go are for instance (adversarial) games that AIs have been built to excel at, courtesy of inordinate effort on the part of AI researchers. So, a question: Prowess in what game by an AI would entail military superiority for the nation that possesses that AI? Answer: Well, contra Scharre, not poker, and more generally, not any game so simple that computational game theory and/or deep learning and/or reinforcement learning can be the basis of an AI's prowess in the game. What's needed is a game that captures real war, in all its hardness. That game, which we here introduce, is Ekte Krig.

To understand Ekte Krig, one must first understand easy games of perfect information, such as Chess, Checkers, and Go. These are all not only Turing-decidable, but — despite what you may have heard from popularizers — of the same particular hardness: they are all in the same exact, rather humble category in the Polynomial Hierarchy, viz. EXPTIME-complete. We next move to another category of easy games, ones whose play with full information is Turing-decidable, but with only partial information get trickier. Our paradigmatic example here will be poker. It is then shown that real war far, far exceeds the simplicity of poker, in any form. This is shown by placing real war [which includes e.g. both espionage and economic strategy in line with (Bringsjord et al. 2012)] within the hierarchies of much harder problems than those in the Polynomial Hierarchy: viz. the Arithmetical and Analytical Hierarchies, and the new Logic Machines hierarchy (LM) that subsumes this pair. After reviewing some Turing-uncomputable perfect-information games from Motalen, with roots going back to (Govindarajulu 2013), we introduce the game of Ekte Krig, Norwegian for “Real War,” and explain: (a) why, unlike Poker, it is logico-mathematically faithful to the hardness of real war, which is easily proved Turing-uncomputable; (b) where, minimally, Ekte Krig falls in LM; and (c) how an AI able to play Ekte Krig can be designed and engineered. Such an AI would be highly destabilizing, because a nation that possesses it would have nonpareil military power. Finally, we explain that whereas ethical correctness of a Poker-playing AI is easily obtained, ensuring ethical correctness of an Ekte Krig-playing AI will be extraordinarily difficult.

### **References**

- Bringsjord, S., Sundar G.N., Eberbach, E. & Yang, Y. (2012) “Perhaps the Rigorous Modeling of Economic Phenomena Requires Hypercomputation” *International Journal of Unconventional Computing* 8.1: 3–32. Preprint available at [http://kryten.mm.rpi.edu/SB\\_NSJ\\_EE\\_YY\\_28-9-2010.pdf](http://kryten.mm.rpi.edu/SB_NSJ_EE_YY_28-9-2010.pdf).
- Govindarajulu, N.S. (2013) *Uncomputable Games: Games for Crowdsourcing Formal Reasoning*. PhD Thesis, RPI.
- Scharre, P. (2023) *Four Battlegrounds: Power in the Age of Artificial Intelligence* (New York, NY: W.W. Norton)

## **HOW DO YOU LIP READ A ROBOT-MITIGATING THE RISKS TRIGGERED BY THE PROCUREMENT OF AI POWERED HR TECHNOLOGY**

SUSAN SCOTT-PARKER

*Business Disability International*

How do you lip read a robot-mitigating the risks triggered by the procurement of AI powered HR Technology; Why do leaders of the global ethical AI debate disregard the potential harm to more than 1.3 billion people living today with disabilities and to the hundreds of millions of us who will become disabled in time? Will the HR cost savings generated by AI technology outweigh the potential damage to the life chances of so many? How will surveillance technology take into account the employer's need, and often legal obligation, to make accommodations for employees who because of a disability do their jobs, and interact with technology, in 'non-standard' ways? Is it true that it is the employer – not the developer - who is liable should a disabled person be treated unfairly because the employer relied on AI Technology? How do you sue an Algorithm? Can we agree on the 5 questions every responsible procurement and HR director should ask every potential supplier, starting with: "Where is your evidence that this 'AI tool' will not discriminate against candidates or colleagues with disabilities?"

### **AI-powered HR technology has a disability problem**

AI recruitment tools have become the first line of defence against high-volume online hiring. But unless the unintended consequences of AI-powered HR technology are urgently addressed, hundreds of millions of people worldwide face lifetimes of economic and societal exclusion.

Just imagine:

- You lose your dream job because your stammer caused you to go 15 seconds over the 3 minutes permitted for the video interview - and the algorithm automatically discards your application.
- You have a facial disfigurement from an acid burn, but the AI tool doesn't recognise your face as real.
- Your visual impairment makes eye contact tricky, but you cannot find any way to request that the video assessment disregard the way your eyes 'dance'.
- Your ADHD means you 'fidget' in front of your screen and the surveillance tech assumes you aren't working –or accuses you of cheating.
- You usually lip-read at interviews, but this robot interviewer is lipless.
- You have used a wheelchair since you were four, but the virtual reality test drops you walking into an ancient tomb to assess your problem-solving skills. You struggle to even imagine standing up (!), never mind doing so while solving complex puzzles.
- And how will you know if your personality profile, produced by scanning everything you have ever put online, tells the recruiter you belong to a Parkinson's Disease self-help network? Is that why your application got nowhere? And if you could take someone to court, who would it be?

AI recruitment tools have become the first line of defence against high-volume online hiring. A recruiter's priority is to discard as many applicants as possible, as quickly and as cheaply as possible, to narrow down to the talent deemed worthy of consideration by human beings. And an increasingly controversial multi-billion-dollar industry stands ready to help.

Thankfully, those influencing responsible AI have begun to address race and gender bias, but the world's 1.3 billion people with disabilities are still so excluded from this debate that no one has even noticed they aren't there.

Brilliantly presented [research by BR](#), the German Public Broadcaster, reveals that a candidate's Behavioural Personality Profile, produced after a one-minute Retorio video interview, changed significantly depending on her appearance. She lost 10 points just by putting on glasses; she gained 20 points by putting on a head scarf. (Retorio says German recruiters find head scarves appealing – so her scores went up).

And what if that camera was to spot your hearing aid, wheelchair, or arthritic hands? Would you score as more or less agreeable, neurotic, or conscientious? No one knows - and that is a problem. [BR didn't test for 'disability indicators'](#). Neither has the developer, nor the German corporations that use it.

Thankfully, those influencing responsible AI have begun to address race and gender bias, but the world's 1.3 billion people with disabilities are still so excluded from this debate that no one has even noticed they aren't there. Unless the unintended consequences of AI-powered HR technology are urgently addressed, hundreds of millions of people worldwide face lifetimes of economic and societal exclusion.

*Neither the AI creators nor their HR customers seem to understand disability discrimination*

Creators often claim they have removed human bias by dropping their AI tool into a standardised recruitment process that treats everyone the same. However, standard processes are by definition inherently discriminatory – recruiters are obliged to make reasonable adjustments at every stage of the process if they want to employ disabled people fairly and on an equal basis. We treat people differently to treat them fairly.

This is not just about the data which, let's face it, is always 'disability biased'. However, biased data, while deeply problematic, is different from the concrete reality of associated discriminatory behaviours, policies, and procedures, such as refusing to adapt an automated talent acquisition process so that a job seeker with a disability can be accurately assessed. And as [recent research from NYU](#) has pointed out, too often the science shaping these assessments is not, in fact, valid for anyone.

AI Creators are not legally obliged to prove their products are 'safe' for any disadvantaged job seekers. But regulators are catching up.

What we have here is a 'market failure'- neither the HR buyers, nor their tech suppliers understand disability discrimination: neither party seems to know how to design a recruitment process that is both barrier-free for people with similar access needs (e.g., accessible game controls) and flexible enough for individuals who need things to be done differently so they can demonstrate their potential (e.g., bypassing psychometric tests which are not validated for people with autism when assessing autistic candidates).

AI creators are not legally obliged to prove their products are 'safe' for any disadvantaged job seekers. But regulators are catching up. The U.S. Equal Employment Opportunity Commission (EEOC) has published their [first guidance](#) which, while still exploratory in nature, does indicate that AI-facilitated disability discrimination is now on its agenda. And the [European Disability Forum](#) is hoping that emerging EU guidance and standards regarding AI accountability will also protect the [human rights](#) of people with disabilities.

Interestingly, a leading HR tech developer, HireView, recently argued that it is the employer that should be held responsible if a candidate claims discrimination further to HireView data being used to justify the decision not to hire them. It's not every day that a

supplier sends such a ‘Buyers Beware’ alert to potential customers. Were employers to respond by requiring all their HR tech service suppliers to prove they have taken the [necessary steps to mitigate potential harm](#) to disadvantaged job-seekers and the associated legal and reputation risks to their brands, it could serve as a compelling reminder of the impact of AI-powered HR technology on disability discrimination.

However, the seriously big question remains: How do we bring the human rights of persons with disabilities into the world view of those influencing this global ethical AI debate? There is no easy answer.

But surely an important first step is to stop the unhelpful waffle about ‘inclusion’ and bring the conversation back to some ‘Disability Equality Basics’. We need a much broader consensus that equality and inclusion are not possible...

- when you can't ask for an interview to be extended because you have a slight speech impairment
- when you aren't told how the employer plans to assess you and therefore cannot ask for the accommodations you require
- when you can't complete the application form using a screen reader
- when you have an intellectual disability and can do the job, but the automated system can't and won't simplify the wording of the interview questions
- when the employer insists you take psychometric tests that have not been validated for sign language users speaking English as a second language
- • when the AI CV screening tool discards your application because it has never heard of [Loyola College](#).

Disability – intrinsic as it is to the human condition – ‘matters’ perhaps more than ever in the age of AI. We need to start using our imaginations and challenge AI creators to develop tools designed explicitly to protect the world’s 1.3 billion people with disabilities from the use of AI-powered HR technology: it’s past time for ‘poachers to turn game keepers’.

## **BENCHMARKING, ROBOTS, AND SMART CITIES**

MATTHEW STUDLEY

*Bristol Robotics Laboratory, University of the West of England, UK*

Robots and Smart Cities are similar in many ways, and have numerous potential synergies which may also expose ethical, legal and societal hazards. Because most people will live in cities and urban environments determine health and well-being to a great extent, these interactions are especially salient. A series of projects and competitions have attempted to benchmark some of these interactions, and this effort continues in an independently funded initiative. This talk will explore some of the issues and opportunities.



**SECTION-2**  
**REGULAR PAPERS**





## EMBEDDING PUBLIC VALUES IN AN ONLINE PUBLIC COMMUNITY NETWORK: A SCENARIO-BASED APPROACH

MATHILDE SANDERS AND ERNA RUIJER

*Utrecht University, Bijlhouwerstraat 6, 3511ZC Utrecht, The Netherlands*  
*E-mail: m.sanders@uu.nl, H.J.M.Ruijer@uu.nl,*  
*www.uu.nl*

JOSÉ VAN DIJCK

*Utrecht University, Achter de Dom 20, 3512 JP Utrecht, The Netherlands*  
*E-mail: j.f.t.m.vandijck@uu.nl*

Public organizations are under pressure to transform many of their activities into digital services (Wirtz & Daiser, 2017; Dunleavy et al., 2006). They need to find out how to transform their public services in the digital context and how these can help create public value (Wirtz et al., 2021; Pang et al., 2014; Panagiotopoulos et al., 2019; Twizeyimana & Andersson, 2019). There is, however, a lack of theoretical clarity on what ‘public value’ means and how digital technologies can contribute to its creation. As most articles on public value creation and e-government are of a conceptual nature, empirical and action-oriented research is welcome (Panagiotopoulos et al., 2019). We use the public business model lens for an empirical study in which we explore individual needs of citizens or users in a future public Decentralized Online Social Network (DOSN).

### 1. Theoretical framework

A public business model outlines how public services create additional value for society and how institutions develop, manage and deliver their services to the public (Wirtz et al., 2021). The focus on customer preferences is one of the central aspects of business models (Foss & Saebi, 2017). The external sphere of the business model is about the formulation of a value proposition and its delivery (Wirtz et al., 2023).

These are key factors for both private and public business models (Mhadevan et al., 2017). A value proposition should answer questions such as: what value is delivered to customers or citizens and what problems are solved for them (Osterwalder & Pigneur, 2010)? The process of how this value proposition is delivered to customer groups (via distribution channels and customer relations) is also an important part of a business model. Public organizations can embed public values into this delivery process.

Public value creation is about the outcome of a process. What our business model and scenario-based design thinking approach adds, is that we illustrate how public value can be delivered by finding ways to incorporate values into the value delivery process (De Graaf, Huberts & Smulders, 2016; De Graaf & Van der Wal, 2010).

All business model frameworks for the public sector in the literature emphasize the importance of integrating user needs into the business model. Astonishingly, however, none of the approaches considers this process (Wirtz et al., 2021). We aim to address these research gaps by answering the following research question: How can public values be embedded in a public decentralized open-source network (DOSN)?

We explore how public organizations can embed public values in their governance of a decentralized online network (DOSN). More specifically, we explore the elements of the value proposition that can be offered to individual users of services offered by public organizations in a future decentralized open source online network (DOSN). We explore which (shared) public

values need to be embedded primarily in the DOSN of a public organization according to its stakeholders. We also analyze what problems public organizations currently encounter regarding public values and how these problems can be solved for the user community of the public organization. These solutions are how public value can be embedded in the software, moderation and organizational governance practices.

## **2. Method**

To answer our research question we deploy both the collective intelligence (Warfield, 2006) and scenario-based design (Carroll, 2000) methods. Via a scenario-based design we collect context-specific user needs and requirements (Ruijter et al. 2017).

Collective intelligence methods gather input from a diverse range of representative stakeholders in the design process and ensure that scenario-based design thinking, incorporating stories about people and their activities (Carroll, 2000), is grounded in a comprehensive understanding of the societal issue. The advantage of scenario-based design is that it is helpful in dealing with complex problems in which the actors have diverging knowledge and backgrounds (Broome, 2017, Janssen et al., 2012, Warfield and Cárdenas, 2002).

We collected data during three workshops held at four Dutch public organizations between November 2022 and March 2023. Two patient associations, one library and one public broadcaster participated. The workshop consists of eight steps of two consecutive rounds of divergence and convergence (individual silent writing, sub-group discussion, and full group presentation and voting).

Based on different user scenarios the workshop participants first identify which public values are most under pressure in the online environment of their public organization. Second, they identify concrete problems connected to these values for their organizations and its user communities. Third, we explore solutions (organizational, software design and moderation) to embed public values in the online public community environment and we identify concrete user needs and digital requirements for embedding these values.

## **3. Findings**

First, we find that privacy, security, user-friendliness and inclusion are the four shared most urgent public values that are currently under pressure for the online communities of the public organizations we studied. Second, we provide an overview of the types of shared concrete problems connected to these (and other) public values that need to be addressed in the governance design of a new public online community network (DOSN). Third, we provide an overview of possible solutions to these problems in the form of technical user requirements, moderation or organizational arrangements.

## **4. Contributions**

First, our study contributes to the literature on public sector business models with a collection of data on concrete user needs, that is currently lacking (Wirtz et al., 2020). Through the co-creation of user scenario's we identify the main target customer groups of public organizations i.e. the citizens that use digital services of public organizations, and their individual needs with regard to the delivery process of public value.

Second, we contribute by connecting the two literatures on public value creation and public values with our business model lens. Public value creation is about the outcome of a process. What our business model and scenario-based design thinking adds, is that we illustrate how public value can be delivered by finding ways to incorporate values into the value delivery

process (via embedding values in moderation, organization and digital governance practices) (De Graaf, Huberts & Smulders, 2016; De Graaf & Van der Wal, 2010). The practical implications of our study are that it provides guidance for managers and public administrators on how to improve public value creation through novel business models (Wirtz et al., 2023).

## References

1. B.J., Broome, Mediating peacebuilding in protracted conflicts: An interactive design framework. I *The Mediation Handbook* (pp. 379-387). Routledge(2017).
2. J.M.,Caroll, "Five Reasons for Scenario-Based Design" in: Introduction to this special issue on 'scenario-based system development'. *Interacting with Computers*, (13), 43-60 (2000).
3. P. Dunleavy, H., Margetts, S. Bastowand J. Tinkler, J. New public management is dead—long live digital-era governance. *Journal of public administration research and theory*, 16(3), 467-494 (2006).
4. G. de Graaf,L. Huberts and R. Smulders Coping with public value conflicts. *Administration & society*, 48(9), 1101-1127(2016).
5. G. de Graaf and Z. van der Wal, Managing conflicting public values: Governing with integrity and effectiveness. *The American Review of Public Administration*, 40(6), 623-630 (2010).
6. N.J. Foss, N. J. and TSaebi,Fifteen years of research on business model innovation: How far have we come, and where should we go?. *Journal of management*, 43(1), 200-227 (2017).
7. M.Janssen, Y. Charalabidis and A. ZuiderwijkBenefits, adoption barriers and myths of open data and open government. *Information systems management*, 29(4), 258-268 (2012).
8. B. Mahadevan,A framework for business model innovation. In *IMRC Conference*, Bangalore (pp. 16-18)(2004).
9. A. Osterwalder and Y. Pigneur, *Business model generation: a handbook for visionaries, game changers, and challengers* (Vol. 1). John Wiley & Sons(2010).
10. P. Panagiotopoulos, B. Klievink and A. Cordella, Public value creation in digital government. *Government Information Quarterly*, 36(4), 101421 (2019).
11. M.S. Pang, G. Lee and W.H.DeLone, IT resources, organizational capabilities, and value creation in public-sector organizations: a public-value management perspective. *Journal of Information Technology*, 29(3), 187-205 (2014).
12. E. Ruijter, S.Grimmelikhuijsen, M. Hogan, S. Enzerink, A.Ojo and A. Meijer Connecting societal issues, users and data. Scenario-based design of open data platforms. *Government Information Quarterly*, 34(3), 470-480 (2017).
13. J.D. Twizeyimana and A. Andersson,The public value of E-Government—A literature review. *Government information quarterly*, 36(2), 167-178 (2019).
14. J.N. Warfield, *An introduction to systems science*. World scientific (2006).
15. J.N. Warfield and A.R. Cárdenas *A handbook of interactive management* (p. 338). Ames: Iowa State University Press (1994).
16. B. Wirtz and P.Daiser,. Business model innovation: An integrative conceptual framework. *Journal of Business Models*, 5(1) (2017)
17. B.W. Wirtz, P.F.Langer and F.W. Schmidt, Digital government: Business model development for public value creation-A dynamic capabilities based framework. *Public Administration Quarterly*, 45(3), 232-255 (2021).
18. B.W. Wirtz, P.R.Kubin and J.C. WeyererBusiness model innovation in the public sector: an integrative framework. *Public Management Review*, 25(2), 3 (2023).

## **INTRODUCING THE CONCEPT OF REPURPOSING ROBOTS; TO INCREASE THEIR USEFUL LIFE, REDUCE WASTE, AND IMPROVE SUSTAINABILITY IN THE ROBOTICS INDUSTRY**

HELEN MCGLOIN

*FARSCOPE CDT, University of Bristol and University of West England,  
Bristol Robotics Laboratory, University of West England, Bristol, UK  
E-mail: [h.mcgloin@bristol.ac.uk](mailto:h.mcgloin@bristol.ac.uk)*

MATTHEW STUDLEY, RICHARD MAWLE, ALAN WINFIELD

*College of Arts, Technology and Environment,  
University of West England, Bristol, UK  
E-mail: [Matthew2.Studley@uwe.ac.uk](mailto:Matthew2.Studley@uwe.ac.uk), [Richard2.Mawle@uwe.ac.uk](mailto:Richard2.Mawle@uwe.ac.uk),  
[Alan.Winfield@uwe.ac.uk](mailto:Alan.Winfield@uwe.ac.uk)*

Based on current definitions of electronic waste, the robotics industry faces a future where products created for both business and consumer markets could be required to meet regulations to manage the control of electronic products at the end of their primary life (e-waste). This paper proposes the new concept of repurposing robots; where a robot at the end of its primary life is repurposed into a new, secondary role, in a process which is generally independent of the Original Equipment Manufacturer [OEM]. By repurposing robots, future waste streams are reduced and the sustainability of the industry is increased. Outlining this new area of work, the authors highlight potential challenges to repurposing which are summarised as topics for future investigation.

### **1. INTRODUCTION**

As electronic waste builds up across the globe, the robotics industry must take responsibility for the current and future waste streams created by the production of robots for business and consumer markets. The open-loop system of designing, producing, using and then discarding electronic products is not sustainable. An open-loop consumer culture harms not only people but the planet around us<sup>1</sup> and can be linked to depletion of natural resources, deforestation,<sup>2</sup> destruction of animal habitats, and pollution of the environment, whether through material leaching or breakdown of plastics into microplastics.<sup>3,4</sup>

While there are many examples of robots being used and developed for the management of other product waste within the recycling and disposal industries,<sup>5-10</sup> little has been written about what happens to robots and autonomous systems at the end of their useful life, when they themselves become product waste. Papers on this topic include Nguyen and Seibel<sup>11</sup> who present the mechanical properties of soft robotic actuators manufactured via recycling of other soft robot actuators, and Steinhilper et al.<sup>12</sup> who demonstrate the application of Steinhilper's remanufacturing process<sup>13</sup> to upcycle a handheld terminal for an industrial robot. This minimal available content reflects the relative infancy of the industry and the limited robotic products currently in circulation. However, as trends in robot number increase, the levels of research and business resource spent on the topic of sustainability must increase.

When making decisions regarding product sustainability, consumers themselves (whether businesses or individuals) are unlikely, when left to their own devices, to make a significant impact in making sustainable choices.<sup>14</sup> Instead, Original Equipment Manufacturers [OEMs] should aim to minimise the environmental impact of their products as part of a commitment to Responsible Innovation.<sup>15,16</sup> Currently, the most accessible method to do this is to recover products at the end of their life by recycling the products at a material level. However,

recycling is still very environmentally wasteful, with useful systems being broken down into individual components and materials. We only have to look at the commonly used Three Rs – Reduce, Reuse, Recycle<sup>17</sup> to see that this hierarchy places Reuse above Recycling.

This paper firstly argues for the need to create a closed-loop system for the robotics industry, based on lessons learnt from the management of other e-waste (Section 2). Secondly, it outlines the options for the management of robotic systems as waste products themselves in Section 3, including presenting the new area of investigation - repurposing robots. Lastly, it presents the challenges of repurposing robotic systems and areas for future interrogation (Section 4).

## 2. THE GROWING PILE OF E-WASTE

### 2.1. *Current e-waste levels*

Products are considered Waste Electrical & Electronic [WEEE] or e-waste when they have reached the end of their useful life and are discarded in a manner where they will not be reused.<sup>18–20</sup> While definitions vary, Electronic & Electrical Equipment [EEE] is described, in general, as any product requiring electrical current or electromagnetic fields to meet its functional purpose.<sup>21</sup> E-waste covers a wide range of products, and examples include: PV panels, professional and household heating systems, washing machines, TVs, printers, photocopiers, mobile phones, toys, non-implanted medical equipment and automated dispensers.<sup>18,22</sup> It includes all “components, sub-assemblies and consumables which are part of the product at the time of discarding”.<sup>21</sup> These lists do not currently include robots and autonomous systems<sup>18,21</sup> and so are not in the scope of current regulations.

Not only are current and historic e-waste levels already worryingly high,<sup>18,19,22–25</sup> but levels of e-waste are also predicted to continue to increase. Savage et al.<sup>26</sup> and Babu et al.<sup>22</sup> predict between three and five percent increase in e-waste year on year within the EU; while Sthiannopkao and Wong<sup>19</sup> predict an increased annual rate between five and ten percent. In 2019 alone, 53.6 million metric tons [Mt] of e-waste were produced globally and predictions by<sup>18</sup> expected this to rise to 74.7Mt by 2030. This is the equivalent of 7.3kgs of e-waste globally produced per capita in 2019 and 9.0kgs in 2030, though levels vary significantly by continent (see Figure 1).

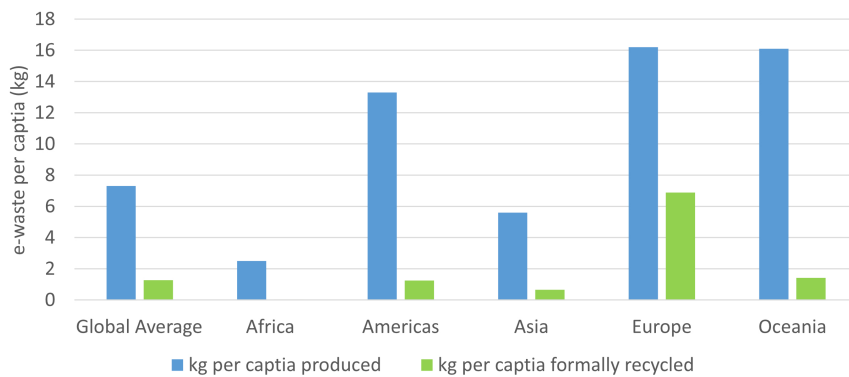


Fig. 1. Global levels of e-waste produced and recycled through formally managed waste systems in 2019<sup>18</sup>

### 2.2. *Why recycling is not good enough*

The United Nations University and collaborators present in their publication The Global E-Waste Monitor 2020<sup>18</sup> four different routes through which e-waste is collected and managed.

These routes are summarised in Figure 2. Only scenario 1 presented in Figure 2 shows the correct management of e-waste. Across the globe, only 40 percent of countries have some form of legislation relating to the proper management of e-waste (scenario 1, Figure 2). However, even with legislation in place, this fails to guarantee the correct management of waste. In the EU, where recycling rates are the highest of any continent (see Figure 1) only 42.5 percent of e-waste follows this route, while 8 percent is discarded directly into municipal waste. Overall, only 17.4 percent of the world’s annual production of WEEE is recycled.<sup>18</sup>

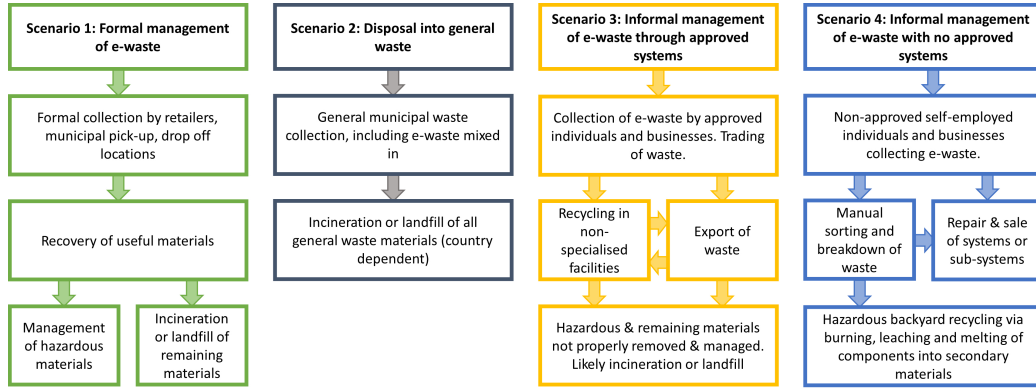


Fig. 2. The four waste management scenarios presented in The Global E-Waste Monitor 2020<sup>18</sup>

In scenarios 2, 3 and 4, inadequate management of WEEE results in the possibility of harmful substances coming into contact with humans and the environment. E-waste sent directly to landfill accounts for 40 percent of lead and 70 percent of heavy metals found in those locations.<sup>22</sup> In addition, large amounts of waste end up via exports in developing countries where approved management systems are either not in place or not overseen.<sup>18</sup> This lack of management results in local populations using waste salvaging as a source of income. Waste salvaging, or ‘backyard recycling’, puts people in direct contact with harmful substances through waste being burnt, leached and melted into resalable materials.<sup>18,19,22</sup>

Knowing that so little of the e-waste intended for recycling actually makes it through controlled waste management schemes, businesses need to consider alternative end-of-life options for e-waste in order to meet future sustainability expectations and requirements.

### 2.3. When robots become e-waste

Robots and autonomous systems are not currently included in the lists of in-scope products for EEE and WEEE legislation. However a robot, by definition, is a “programmed actuated mechanism with a degree of autonomy to perform locomotion, manipulation or positioning”.<sup>27</sup> Where the power mechanism used to provide locomotion, manipulation or positioning to the system is electrical, it can be concluded that robots meet the general definitions of EEE.

At the time of publishing of the EU Directive in 2002, most robotic systems were generally limited to industrial settings where they could be classed within the ‘large stationary tools’ category, which is exempt from the directive.<sup>21</sup> However, as technology develops, the use of robots is evolving from stationary manufacturing roles into service-based roles,<sup>28</sup> which whether used in the home or in a professional setting, could be considered to be e-waste when they cease to be required or able to perform their original function.

In addition, the rate of growth of robotic products entering the consumer and business markets will likely result in greater scrutiny of robots as waste products when they reach the

end of their life. Taking only the service sector (including service robots in professional and domestic settings<sup>29</sup>), the International Federation of Robotics [IFR] reported a 12 percent increase in the service robot market in 2020 and also noted that 17 percent of the robot suppliers surveyed in this sector were start-ups.<sup>30</sup> The report observed that the majority of the start-ups, and a large number of the established businesses, had products still in the development stage that were not available for commercial use.<sup>30</sup> This growth in the robotics market is further supported by market analysts, Mordor Intelligence LLP, who predict the domestic robot market globally will increase from 6.8bn USD in 2021 to 21.9bn USD by 2027.<sup>31</sup>

Businesses and research groups working within the robotics industry, therefore, have the opportunity to anticipate the inclusion of robotic products in future e-waste definitions ahead of any potentially related legislation. In addition, opportunities for designing for a circular economy will be easier to introduce during the early stages of a product development life-cycle, rather than retrofitting to meet requirements. The book ‘Designing for the Circular Economy’ notes that 80 percent of a product’s environmental impact is determined at the early stages of product development.<sup>32</sup>

The unique characteristics of a robotic system over other non-robotic electronic products mean that, beyond the option for recycling or the reuse of all or part of the system, robots have the potential to be repurposed into new, secondary roles. These secondary uses can be significantly different from their original primary function or design, though the resulting product is still classed as a robot. This is in comparison to other electronics such as mobile phones, or mechanical-electrical products such as cars. Both phones and cars can be recycled for parts, or they can be reused in their current state and resold in the second-hand market.

### 3. REPURPOSING AND THE ALTERNATIVE Rs

#### 3.1. *Definition for repurposing a robot*

The authors propose the repurposing of a robot is defined as: **providing new utility to an existing robotic system in order to give the system a new role which is independent of the robot’s original utility.** In this definition, a robot’s utility is comprised of:

- Skill - the tasks which the robotic system is capable of completing, and
- Application - the context in which the robotic system is capable of functioning.

Therefore, both the skill and application of the robot must be changed in order to meet the definition of repurposing. This is supported by the British Standards Institute’s general definition of repurposing, where a product or component is utilised ‘in a role that it was not originally designed to perform’.<sup>20</sup> In order to change the skill and application of the robotic system, resources in the form of time and/or cost must be applied.

Using this definition, an example of repurposing would be to take an industrial robot that had previously been used in a production line and repurpose it into a robot being utilised in a hospital setting collecting waste. In this example, in its primary life, the robotic arm would have been required to place fixtures into a product at high speeds and high accuracy. Once its performance levels could no longer be maintained, the robot would be considered to be at the end of its primary life. By integrating the original robotic system onto a mobile base, adding vision systems and improving safety protocols, the robot could be repurposed to collect and sort waste in a hospital setting. This example meets the proposed definition of repurposing as both the skill and the application of the robotic system have changed and it would no longer be able to meet the utility requirements of the original system. This example is further illustrated in stages A and C in Figure 3.

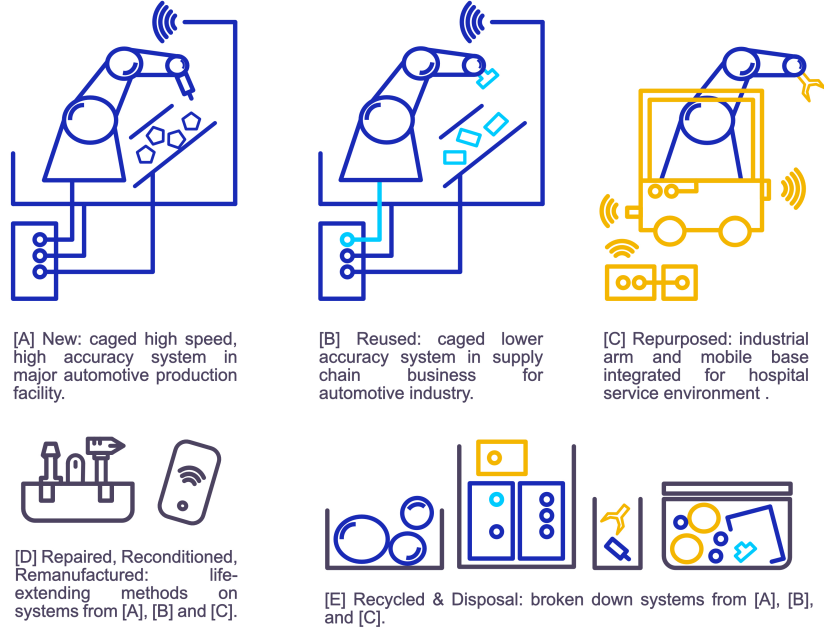


Fig. 3. Example of a robotic system as it goes through the different life conditions

Repurposing in this way is unique to robotic systems in comparison to other electronic devices. A robot, by its definition (Section 2.3), can be totally reprogrammed to change its skill, which enables it to work in a new application, sometimes with the application of additional hardware. This can be compared to the way that humans can re-skill themselves by learning new skills and taking on new tools in order to take on a role in a different sector or industry to the one they originally worked in. Many robots designed for both domestic and industrial applications should also be able to complete this transformation through repurposing. In comparison, a phone or smart washing machine might be re-programmable, but they would need to be broken down into component or sub-system level, and rebuilt into an entirely new system in order to fundamentally change their skill or application.

### 3.2. Alternatives to repurposing

Alongside repurposing, there are a number of alternative management routes which can be used for robotic waste at the end of its primary life; recycling and disposal; repair, remanufacturing and reconditioning; and reuse (grouped as similar processes as described in Figure 4). These processes, are far better established than repurposing and can be described as follows:

- Repair - where work completed by either the product owner, OEM or independent service provider returns the product to a working condition following damage or wear.<sup>20</sup> Following the repair, the product will meet all or most of the original utility requirements (skill and application) of the robot, though it is possible for some functionality to be lost over time. However, repairs must bring the system up to an acceptable working level, as set by the customer's requirements. Repair is illustrated by stage D in Figure 3.
- Recondition - an intermediate between repair and remanufacture where, before failure occurs, components or subsystems are repaired, returning the product to a good working condition.<sup>20</sup> Like repair, there may be some functionality performance loss



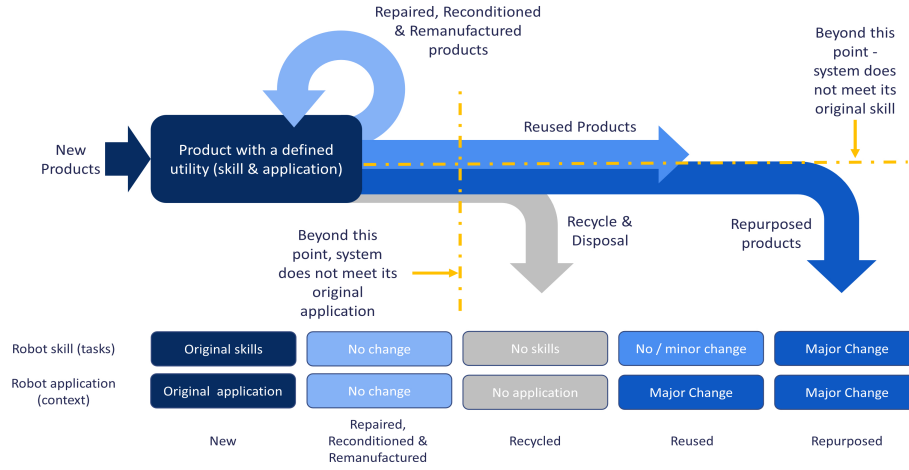


Fig. 4. A flowchart of waste management routes for robots at the end of their primary life with a comparison of skills and application.

to the original specification. Reconditioning can also be referred to as refurbishing<sup>20</sup> and for the purpose of this paper will be included within the concept of repair. Reconditioning is illustrated by stage D in Figure 3.

- Remanufacture - where a product is split and disassembled to component level; and components are then inspected, tested, and repaired or replaced if necessary, before reassembling.<sup>12,13,33</sup> In this way, the robotic product is up-cycled, with the resulting product meeting, as a minimum, the utility condition of the original product specification. During this process, it can be useful to upgrade beyond the original specification to the new accepted standard - whether this is hardware or software improvements. Remanufacture is illustrated by stage D in Figure 3.
- Reuse - where a robotic product of a given utility is placed into a new application (context). This may require minor changes to both hardware and software of the original product in order to meet the new, secondary, role. Though generally, the robot will still meet the skill functionality of the original design. A product that has been reused is capable of returning to its original application with only minor changes to its hardware or software. Reuse is illustrated by stage B in Figure 3.
- Recycle and disposal - where both working and irreparable products are disposed of either into landfill or to be recycled. During the recycling process, products are broken down into useful parts and materials for reuse at the material or component level. Products that are sent for recycling and disposal are no longer able to meet their original utility conditions following the completion of this process. Recycling and disposal are illustrated by stage E in Figure 3

Figure 4 demonstrates the flow of robotic products as they reach the end of their primary life, based on the selected waste management route. For each route, the change in skill and application in comparison to the original utility is given. In addition, Figure 3 provides an example of how each management route could be applied to an industrial robot arm.

Additionally, it should be noted individuals and businesses may retain or store their old robotic systems at the end of their useful life. The habit of retention of old tech arises from a perception that old electronics have intrinsic value that may be realised at a later date<sup>22</sup> and is commonly referred to as hibernation - the difference between the total time an electronic product is owned and the total time it is used.<sup>34,35</sup> The United States Environmental Protection Agency calculated in 2020 that 70 percent of consumer waste electronics

were stored for a period of three to five years following the end of their use.<sup>22</sup> This figure does not currently include consumer robots. However, should robots also be hibernated in this way at the end of their primary life, their utility would be negatively affected; retaining their skill value but losing all application. At any point, the hibernated robotic system could be repaired, refurbished, remanufactured, reused, repurposed or recycled, thereby giving it a new life. For this reason, stored robotic systems are not included in their own category within the scope of the defined management routes for robots at the end of their primary life and are not included in Figure 4 or Figure 3.

#### 4. CHALLENGES TO REPURPOSING

Of the routes available to businesses to manage a robot at the end of its primary life, repair and reconditioning are the most accessible options as both OEM's and independent service providers are generally able to carry out repair and reconditioning work. In comparison, remanufacturing has yet to become well-established in the robotics industry. However, as more robots reach the end of their primary life, it is likely that remanufacturing will become a more commercially attractive option for those who wish to maintain value in a robot investment without the outlay for new systems. Alternatively, should a robot reach the end of its primary life and the robot owner no longer requires the system but still considers it has available utility, reuse is an already established option. Individuals are able to purchase used robotic systems through resale either directly with the owner or via brokers and auctioneers. The new owner must invest in the system in order to make minor changes to the utility of the robot, while the original owner sees a partial return on the original purchase costs. Lastly, should repair, reconditioning, remanufacturing or reuse not be possible, then it is currently possible to recycle and/or dispose of the system through verified processes.

As recycling and disposal have been demonstrated to remove all available utility within the robotic system, the resource that was placed into the system as part of its design and manufacture is lost. However, if repurposing can be made a viable and functional option for systems at the end of their primary life, it would further delay the point at which a robotic system would require being broken down into residual components or sub-systems for recycling or disposal, thereby reducing annual waste production levels and increasing product life. Repurposing is an entirely new field of study in the area of robotics. As such, it naturally faces a number of challenges, which must be explored in order to demonstrate successfully the process and the potential of repurposing a robotic system. A number of these challenges will be addressed in future work, and are outlined in Sections 4.1 to 4.3.

##### 4.1. *Viability*

In order to evaluate the suitability of a system for repurposing, a method must be created to assess the viability of repurposing a given system; assessing the ease of repurposing the system versus the expected utility received from completing the repurposing process. This evaluation would produce a repurposability metric. Figure 5 proposes an example relationship between utility (y-axis) and cost (x-axis) for repurposing a robotic system that requires interrogation in future work. Here the utility is shown as a percentage calculated against a like-new system, where the cost of a new system is also known. In scenario A (of Figure 5) the cost of repurposing is lower than the cost of a new system, and the utility of the repurposed system meets an acceptable minimum requirement (threshold). This scenario suggests that the repurposing of the proposed system should go ahead. In scenario B the required minimum utility is not met but the cost of the system is lower than the purchase of a new system. It may be possible in this scenario to review or amend requirements or find an alternative application that will make repurposing viable.

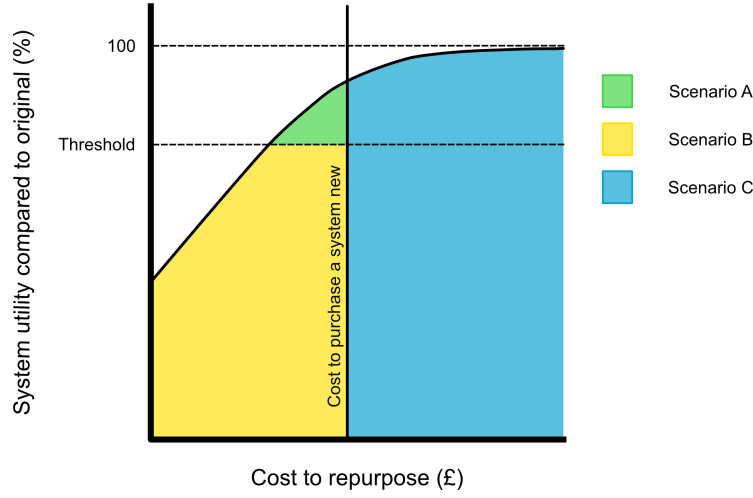


Fig. 5. Proposed relationship between the utility and cost of repurposing a robotic system to be investigated in future work

In scenario C, repurposing costs are higher than those for buying a new system and it would not be economically justifiable to select repurposing. However, repurposing in scenario C could still be justified on sustainability grounds. Assuming a new robotic system and a repurposed system are able to meet the same or acceptably similar utility, the relationship between the sustainability of the system and the cost must be understood. Methods of calculating a sustainability index are required in order to understand the impact of producing a new versus a repurposed system. This should include assessing the carbon footprint to manufacture new systems versus a repurposed system. It should additionally provide an understanding of other environmental factors including, but not exclusive to; resource extraction resulting in deforestation, loss of animal habitats, and chemical leaching.

Future work required on the viability of repurposing must include an investigation into measurable indices for robot utility, cost and sustainability versus new systems. For the cost of repurposing to be calculated, technical feasibility must be understood (Section 4.2).

#### 4.2. Technical

All robotic systems consist of hardware and software. The integration of these systems is generally purpose-built to enable the product to meet its desired functional requirements. As repurposing amends both the skill and application of the robotic system to meet the required new utility, it is likely that both the hardware and software will need amending in order to meet the new requirements. Integration of these amended systems will not be simple. Since this is a new concept, no examples of repurposing a robotic system exist. Therefore, a method must be developed and tested to demonstrate the process, detailing its successes and limitations. Verification of the repurposed systems must be equivalent to the verification processes for new systems, in order for the process to be accepted by customers.

In addition to this, changes in the technical capabilities of robotics, such as opportunities provided by advances in morphological computation, artificial intelligence, an increased Human-Robot-Interaction [HRI], and improvements in standardisation and modular sub-systems for both hardware and software, may affect the technical possibility of repurposing a robot.

For example; improvements in HRI, which will affect both the hardware and software of robots, will have the potential to create greater levels of safety and understanding between robotic systems and human operators and the general public. These robotic systems will

not only be programmed to better cope with a wider variety of human interactions,<sup>36,37</sup> but often will be made of more compliant materials.<sup>38</sup> Both these factors could make the systems more desirable for repurposing. However, by taking a robot out of its original application with the intention of repurposing, challenges will likely appear in the verification of the system to ensure it continues to react in an expected manner if its role type has changed.

Likewise, morphological computation - where aspects of a robot's control, perception or cognition are removed from traditional programming and instead are designed to occur naturally in the robotic body itself to mimic capabilities seen in nature<sup>39</sup> - may produce robotic systems that are more susceptible to being repurposed. Characteristics of robotic systems with morphological capabilities that could be advantageous in repurposing include; increased system flexibility due to high levels of sensing embedded within the robot;<sup>40</sup> increased dexterity for systems that are designed for self-stabilisation;<sup>41</sup> or an ability to control a larger ensemble of otherwise low capability robots to complete tasks that could not be met individually.<sup>42</sup> Note however, the robotic systems currently developed with aspects of morphological computation are often in very specialist forms that are dictated by their need to meet sensing and self-stabilising requirements. They may therefore offer limited repurposing opportunities at the end of their primary life as a result of their physical form.

As the variety in forms and functionality of robots expands, it is realistic to assume that some advancements will support repurposing, while others will hinder the process. Only once the process of repurposing is validated on currently well-understood technologies such as with industrial arms, will it be possible to assess the effects of new capabilities within the robotics industry.

### ***4.3. Attitudes, Incentives and Legislation***

Alongside demonstrations that it is technically possible and viable to repurpose a robot, a market must also exist for the resultant product. It has already been shown that in the consumer market, push, pull and mooring factors influence a potential consumer's decision to switch to remanufactured goods.<sup>43</sup> Mooring factors are determined by the consumers' pre-existing attitudes and cover social and personal values. Key pull factors influencing consumers are; incentives by governments (including legislation, tax and subsidies), and knowledge of the environmental benefits of switching to second-hand goods.<sup>43</sup> The key push factor that drives a consumer's decision to purchase remanufactured goods is the perception or realisation of the comparative savings that may be achieved by comparison with buying new products.<sup>43</sup> Investigation into the attitudes of the general public to second-hand robots compared to new robots for the consumer market will be considered in future work through participant surveys.

It is unlikely that consumer action alone will result in a significant impact in making sustainable choices.<sup>14</sup> Instead, product researchers, designers, engineers and manufacturers should take responsibility during the design phase to implement methods which minimise the environmental impact of their products as part of a commitment to Responsible Innovation.<sup>15</sup> Repurposing will be better achieved if product OEMs themselves take steps to enable the repurposing process of their robotic products. However, it can be expected that the concept of repurposing robots will have a similar reception to that of the smartphone industry to the the Right to Repair Movement. The Right to Repair movement is the global campaign to see improved consumer rights relating to the repair of electronic goods such as mobile phones and washing machines.<sup>44</sup> In general, large technology companies have not been supportive of the introduction of Right to Repair laws and legislation, with vocal opposition coming from Apple, Microsoft, Tesla, Amazon and many others.<sup>45,46</sup> Typical concerns raised by technology companies include; the security of devices being compromised by opening up

to independent repairs, sensitive diagnosis information being made available to competitors, the safety of products if incorrectly repaired, and future infringements of copyright laws as a result of opened access.<sup>45,47–50</sup> Countering this, the Right to Repair movement accuses technology companies of deliberately designing obsolescence and inaccessible repairs in their products, which fuels the high levels of e-waste around the world<sup>51</sup> as shown in Section 2.

Despite their original opposition to the effects of Right to Repair in the electronics industry, changes to laws and regulations (in development and published) in the US, UK and EU have persuaded technology firms to reconsider their stance on repairing consumer products. Examples include Apple Inc who, in November 2021, launched the Self Service Repair option that makes parts, tools and manuals available to individual customers and independent repair shops<sup>52,53</sup> despite their earlier resistance to the concept. Assuming robotics business concerns about repurposing will be similar to the initial reactions to Right to Repair, lessons can be learnt from the progression of the Right to Repair movement.

Understanding the concerns, constraints and benefits to robotic businesses of repurposing will be important to enabling uptake of the concept, should technical and viability studies prove successful. For this reason, a qualitative research study will be carried out that will interview robotics industry experts to understand the perception of robots as e-waste, repair accessibility and attitudes to the concept of repurposing. Interviews will include both businesses that develop or produce robotic systems, and those who use robot products. Data gathered from these studies will be used to address challenges for repurposing in relation to the topics of safety, security, sustainability and legislation.

## 5. CONCLUSIONS

What happens to robots when they have finished their primary use is a topic which, to date, has not been given substantial consideration within the industry. Robots may be utilised in the management of other product waste, but they themselves are rarely thought of as waste products. Although robots are not currently included in the definition lists of e-waste for regulations, they do meet the general definition. With the number of robots increasing in both work and domestic settings, and with a greater global focus on the sustainability of all industries, those who are involved in developing robotic products must expect greater scrutiny of how those systems are disposed of at the end of their primary life.

This paper has argued that recycling should be the last resort when it comes to the management of robotic e-waste due to the examples set by other electronic products, which are subject to poor recycling rates and poor waste management practices in the recycling methods. Instead, the concept of repurposing has been proposed as a method to delay the time before a robotic product is discarded for recycling and disposal. When carried out, the repurposing of a robotic system will increase the useful life of the product and contribute to a circular economy. The authors have defined repurposing as providing new utility to an existing robotic system in order to give the system a new role that is independent of the robot’s original utility, with utility being a combination of skills and application of the robot. As an entirely new field of study, this paper has outlined the major challenges and opportunities for repurposing that will be investigated by the authors in future work.

## Acknowledgments

The work of Helen McGloin was supported by the EPSRC Centre for Doctoral Training in Future Autonomous, and Robotic Systems (FARSCOPE).

## References

1. S. Carlisle and P. Hanlon, *Health Promotion International* **22**, 261(Sept. 2007).
2. P. Dauvergne and J. Lister, *Organization Environment* **23**, 132(Jun. 2010).
3. M. Smith and D. C. Love, *Current environmental health reports* **5**, 375(Sept. 2018).
4. Z. Huang, Y. Weng, Q. Shen, Y. Zhao and Y. Jin, *Science of The Total Environment* **785**, p. 147365(Sept. 2021).
5. C. Spelhaug, Amp robotics installs its first recycling robots in the united kingdom and ireland with recyco AMP Robotics Corp. News [online] [www.amprobotics.com/news-articles/amp-robotics-installs-its-first-recycling-robots-in-the-united-kingdom-and-ireland-with-recyco](http://www.amprobotics.com/news-articles/amp-robotics-installs-its-first-recycling-robots-in-the-united-kingdom-and-ireland-with-recyco) (accessed Aug. 15, 2022), (2021).
6. E. Slow, Viridor invests in max-ai sorting robot letsrecycle.com [online] [www.letsrecycle.com/news/viridor-max-sorting-robot/](http://www.letsrecycle.com/news/viridor-max-sorting-robot/) (accessed Aug. 15 2022)(Aug., 2018).
7. S. Faibish, H. Bacakoglu and A. Goldenberg, *Proceedings - IEEE International Conference on Robotics and Automation* **1**, 9 (1997).
8. L. Chin, J. Lipton, M. C. Yuen, R. Kramer-Bottiglio and D. Rus, *RoboSoft 2019 - 2019 IEEE International Conference on Soft Robotics* , 102(May 2019).
9. J. Li, M. Barwood and S. Rahimifard, *2014 IEEE International Electric Vehicle Conference, IEVC 2014* (Mar. 2015).
10. M. Bounouar, R. Bearee, A. Siadat, N. Klement and T. H. Benchekroun, *2020 1st International Conference on Innovative Research in Applied Science, Engineering and Technology, IRASET 2020* (Apr. 2020).
11. A. Nguyen and A. Seibel, *2022 IEEE 5th International Conference on Soft Robotics, RoboSoft 2022* , 571(Apr. 2022).
12. R. Steinhilper, J. Kleylein-Feuerstein and C. Kussmann, *2016 Electronics Goes Green* (Jan. 2017).
13. R. Steinhilper, *Remanufacturing: the ultimate form of recycling* (Fraunhofer IRB Verlag, 1998).
14. A. Lubowiecki-Vikuk, A. Dabrowska and A. Machnik, *Sustainable Production and Consumption* **25**, 91(Jan. 2021).
15. UKRI, Responsible innovation UK Research and Innovation [online] [www.ukri.org/about-us/policies-standards-and-data/good-research-resource-hub/responsible-innovation/](http://www.ukri.org/about-us/policies-standards-and-data/good-research-resource-hub/responsible-innovation/) (accessed May 9, 2022)(Oct., 2021).
16. A. Winfield and K. Winkle, *ICRES 2020: 5th International Conference on Robot Ethics and Standards* , 28(Sept. 2020).
17. Our waste, our resources: A strategy for england HM Government via gov.uk [online] [assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment\\_data/file/765914/resources-waste-strategy-dec-2018.pdf](https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/765914/resources-waste-strategy-dec-2018.pdf) (accessed Sept. 9, 2022), (2018).
18. V. Forti, C. Baldé, R. Kuehr and G. Bel, The global e-waste monitor 2020: Quantities, flows and the circular economy potential United Nations University / United Nations Institute for Training and Research – co-hosted SCYCLE Programme, International Telecommunication Union International Solid Waste Association, Bonn/Geneva/Rotterdam., (2020).
19. S. Shianopkao, M. H. Wong, R. Naidu and M. Wong, **463-464**, 1147(Aug. 2012).
20. Design for manufacture, assembly, disassembly and end-of-life processing (made) part 2: Terms and definitions BS 8887-2:2009, British Standards Institute, (2009).
21. Directive 2002/96/ec of the european parliament and of the council of 27 january 2003 on waste electrical and electronic equipment (weee) EUR-Lex (access to European Union Law) [online] <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A32002L0096> (accessed May 6, 2022).
22. B. R. Babu, A. K. Parande and C. A. Basha, *Waste Management and Research* **25**, 307(Aug. 2007).
23. B. C. Zoeteman, H. R. Krikke and J. Venselaar, *The International Journal of Advanced Manufacturing Technology* **2009 47:5** **47**, 415(Oct. 2009).
24. R. Widmer, H. Oswald-Krapf, D. Sinha-Khetriwal, M. Schnellmann and H. Böni, *Environmental Impact Assessment Review* **25**, 436(Jul. 2005).
25. ENVIROS, Potential markets for waste electronic and electrical equipment (weee) Funded by the UK Environment Agency(Jun., 2002).
26. M. Savage, J. Lindblom and L. Delgado, Implementation of waste electric and electronic equipment directive in the eu 25 - publications office of the eu Institute for Prospective Technological Studies [online] <https://op.europa.eu/en/publication-detail/-/publication/>

- 9e42f67b-2c1a-4cfa-9a27-3955e0a50856 (accessed May 6, 2022)(Aug., 2007).
27. Robotics — vocabulary ISO 8373:2021(en), International Organization for Standardization [online] [www.iso.org/obp/ui/#iso:std:iso:8373:ed-3:v1:en](http://www.iso.org/obp/ui/#iso:std:iso:8373:ed-3:v1:en) (accessed Aug. 25, 2022), (2021).
  28. V. Graefe and R. Bischoff, *ICEMI 2009 - Proceedings of 9th International Conference on Electronic Measurement and Instruments*, 3418 (2009).
  29. IFR, Chapter 1: Definitions and classifications of service robots [https://ifr.org/img/worldrobotics/Definitions\\_WR\\_Service\\_Robots\\_2021.pdf](https://ifr.org/img/worldrobotics/Definitions_WR_Service_Robots_2021.pdf).
  30. IFR, Executive summary world robotics 2021-service robots International Federation of Robotics [online] [ifr.org/img/worldrobotics/Executive\\_Summary\\_WR\\_Industrial\\_Robots\\_2021.pdf](http://ifr.org/img/worldrobotics/Executive_Summary_WR_Industrial_Robots_2021.pdf) (accessed Feb. 14, 2022), (2021).
  31. Household robots market - growth, trends, covid-19 impact, and forecasts (2022 - 2027) ReportLinker [online] [www.reportlinker.com/p06241302](http://www.reportlinker.com/p06241302) (accessed Jan. 27 2023)(Feb., 2022).
  32. M. Charter (ed.), *Designing for the Circular Economy* (Routledge, 2019).
  33. N. Nasr, *Remanufacturing in the circular economy : operations, engineering and logistics* (John Wiley Sons, Inc. ;, 2019).
  34. S. Murakami, T. Tasaki, I. Diago and S. Hashimoto, *Journal of Industrial Ecology* **14**, 598 (2010).
  35. G. T. Wilson, G. Smalley, J. R. Suckling, D. Lilley, J. Lee and R. Mawle, *Waste Management* **60**, 521(2 2017).
  36. A. Hong, N. Lunscher, T. Hu, Y. Tsuboi, X. Zhang, S. F. dos Reis Alves, G. Nejat and B. Benhabib, *IEEE Transactions on Cybernetics* **51**, 5954(Dec. 2021).
  37. A. Gaschler, S. Jentzsch, M. Giuliani, K. Huth, J. D. Ruiter and A. Knoll, *IEEE International Conference on Intelligent Robots and Systems*, 2128 (2012).
  38. A. Ueno, V. Hlavac, I. Mizuuchi and M. Hoffmann, *29th IEEE International Conference on Robot and Human Interactive Communication, RO-MAN 2020*, 14(Aug. 2020).
  39. V. C. Müller and M. Hoffmann, *Artificial Life* **23**, 1(Feb. 2017).
  40. E. Judd, G. Soter, J. Rossiter and H. Hauscr, *RoboSoft 2019 - 2019 IEEE International Conference on Soft Robotics*, 558(May 2019).
  41. R. Terajima, K. Inoue, S. Yonekura, K. Nakajima and Y. Kuniyoshi, *IEEE Robotics and Automation Letters* **7**, 1597(Apr. 2022).
  42. R. Warkentin, W. Savoie and D. I. Goldman, *Proceedings - 2nd IEEE International Conference on Robotic Computing, IRC 2018*, 224(Apr. 2018).
  43. B. Hazen, D. Mollenkopf and Y. Wang, *Business Strategy and the Environment* **26**, 451(May 2017).
  44. N. Šajn, Right to repair briefing European Parliamentary Research Service(Jan., 2022).
  45. R. Mittal, Right to repair movement around the world ipleaders [online] [blog.ipleaders.in/right-repair-movement-around-world/](http://blog.ipleaders.in/right-repair-movement-around-world/) (accessed May 5, 2022), (2021).
  46. J. Vipra and S. Rao, 'right to repair', the legislation india needs to save money, minimize e-waste The Federal [online] [thefederal.com/analysis/upgrades-electronic-device/](http://thefederal.com/analysis/upgrades-electronic-device/) (accessed May 9, 2022), (2021).
  47. C. Godwin, Right to repair movement gains power in us and europe BBC News [online] [www.bbc.co.uk/news/technology-57744091](http://www.bbc.co.uk/news/technology-57744091) (accessed May 5, 2022)(Jul., 2021).
  48. P. Arora, Does india need a 'right to repair' legislation? The Daily Guardian [online] [thedailyguardian.com/does-india-need-a-right-to-repair-legislation/t](http://thedailyguardian.com/does-india-need-a-right-to-repair-legislation/t) (accessed May 9, 2022), (2021).
  49. M. Bergen, Microsoft and apple wage war on gadget right-to-repair laws Bloomberg [online] [www.bloomberg.com/news/articles/2021-05-20/microsoft-and-apple-wage-war-on-gadget-right-to-repair-laws](http://www.bloomberg.com/news/articles/2021-05-20/microsoft-and-apple-wage-war-on-gadget-right-to-repair-laws) (accessed May 9, 2022), (2021).
  50. L. Dormehl, Tech companies kill right to repair MUO (Make Use of) [online] [www.makeuseof.com/tech-companies-kill-right-to-repair/](http://www.makeuseof.com/tech-companies-kill-right-to-repair/) (accessed May 9, 2022), (2021).
  51. Nixing the fix: An ftc report to congress on repair restrictions Federal Trade Commission [online] [www.ftc.gov/system/files/documents/reports/nixing-fix-ftc-report-congress-repair-restrictions/nixing\\_the\\_fix\\_report\\_final\\_5521\\_630pm-508\\_002.pdf](http://www.ftc.gov/system/files/documents/reports/nixing-fix-ftc-report-congress-repair-restrictions/nixing_the_fix_report_final_5521_630pm-508_002.pdf) (accessed May 09, 2022(May, 2021).
  52. Apple announces self service repair - press release Apple Inc. Newsroom [online] [www.apple.com/newsroom/2021/11/apple-announces-self-service-repair/](http://www.apple.com/newsroom/2021/11/apple-announces-self-service-repair/) (accessed May 9, 2022), (2021).
  53. D. Schneider, Did apple really embrace right to repair Business of Apps [online] [spectrum.ieee.org/right-to-repair-apple](http://spectrum.ieee.org/right-to-repair-apple) (accessed May 9, 2022), (2021).

## NON-PROFIT-DRIVEN DEVELOPMENT PATHS FOR EQUITABLE AI/AGI

MICHAEL GIANCOLA<sup>\*</sup> and JAMES OSWALD<sup>†</sup>

*Rensselaer AI & Reasoning (RAIR) Lab<sup>†</sup>*  
*Department of Computer Science<sup>\*†</sup>; Department of Cognitive Science<sup>\*</sup>*  
*Rensselaer Polytechnic Institute, Troy, NY 12180, USA*  
*E-mail: { mike.j.giancola, james.oswald.111 } @gmail.com*

The organizational structure within which AI technologies are developed has a significant effect on the distribution of their benefits. In particular, the extent to which AI technologies developed within for-profit corporations are made available is completely up to the corporation. This poses a serious concern when the prospect of artificial general intelligence (AGI) is considered. Specifically, we see evidence that if AGI is ultimately produced within a for-profit business, it is unlikely to be aligned with the interests of humanity at large. We discuss several alternative development paths for AI/AGI and analyze their potential and limitations with regard to equitably distributing AI technologies. We end by looking at organizations trying to achieve AGI and classifying them with respect to these pathways.

*Keywords:* Equitable AI; ethical AI; artificial general intelligence; not-for-profit business.

### 1. Introduction

“The way technology currently operates is optimized for populations in power rather than everyone affected by it. In the absence of force, some actors are unlikely to make meaningful progress toward equity because they prioritize profits.”

—Aspen Institute Science & Society Program

The organizational structure within which AI technologies are developed has a significant effect on the distribution of their benefits. In particular, the extent to which AI technologies developed within for-profit corporations are made available is completely up to the corporation. Consider the recent release of GPT-4 by OpenAI, which recently transitioned from non-profit status to a capped-profit business model.<sup>a</sup> While they released a 100-page technical report on the model, it was nearly devoid of detailed information regarding the model or datasets used to train the model.<sup>b</sup> In particular, the authors state the following in the report:

Given both the competitive landscape and the safety implications of large-scale models like GPT-4, this report contains no further details about the architecture (including model size), hardware, training compute, dataset construction, training method, or similar. [1]

This example is one of many instances of for-profit companies obscuring the details of their AI models. This lack of transparency poses a serious concern when the prospect of artificial general intelligence (AGI) is considered. Specifically, we see evidence that if AGI

---

<sup>a</sup>We will discuss the capped-profit business model, and OpenAI’s implementation of it, in §3.2.

<sup>b</sup>The authors candidly admit this in the report. Essentially the only information given about the model was that: (1) it is a “Transformer-style model”, (2) it was trained on both public and non-public data, and (3) it was “fine-tuned using Reinforcement Learning from Human Feedback (RLHF).” [1]



is ultimately produced within a for-profit business, it is unlikely to be aligned with the interests of humanity at large.

Montes & Goertzel [2] discuss the potential harms of centralized, for-profit mega-corporations developing AGI. To wit, these corporations are driven primarily by the goal of generating profit, and are unlikely to make decisions in the best interest of humanity at large if such a decision conflicts with their ability to generate adequate profits. Therefore it is critical that those in the AGI community consider the benefits and risks of the various potential development paths for AGI. However, we don't have to wait for the advent of AGI to begin considering and addressing these concerns; there is already reason to be concerned with the closed-door development of AI technologies.

Bender et al. [3] discuss several financial and environmental concerns surrounding the development of large language models. One of the authors of the paper — Timnit Gebru — was the co-lead of Google's ethical AI team at the time the paper was submitted for publication. While Google's internal review process initially approved the paper for submission, the authors were later "told to either retract the paper or remove their names." [4] Gebru pushed back and was subsequently terminated. While Gebru affirms that she was fired, Google claims that they simply accepted her resignation [5]. However, it is, at minimum, *plausible* that her termination was related to the paper, since it criticized large language models, technology which Google was invested in [4].

Martin [4] asked the salient question: "who is responsible for the critical evaluation of the products and services developed by Big Tech?" While we don't propose an answer, we argue one response is certainly incorrect: the company itself. It is evident that for-profit companies cannot be trusted to expound criticism of AI technologies from which they can generate profit.<sup>c</sup>

We argue that AI researchers — and *especially* AGI researchers — should consider and pursue alternative development paths.<sup>d</sup> Potential paths should be evaluated with regard to their ability to align the business structure's goals with those of equitable AI development. In §2, we will describe what we mean by *equitable* AI. Thereafter, in §3, we will present and analyze five potential paths we see as having merit for equitable AI development:

- §3.1: decentralized AI architectures,
- §3.2: the capped-profit business model,
- §3.3: the not-for-profit business model,
- §3.4: an independent agency or government-run program,
- §3.5: academic laboratories.

## 2. Equitable vs. Ethical AI

"It has never been clearer that we share an urgent responsibility to ensure that scientific and technological advances serve the many, and not just the few."

—Aspen Institute Science & Society Program

What do we mean by *equitable* AI? And how does it differ from *ethical* AI? Ethical AI is a fairly broad category; generally speaking, any work in the intersection of ethics and

---

<sup>c</sup>In fact, some argue that the sudden proliferation of corporate "ethical AI" groups was motivated by these companies' desire to forestall the government from regulating controversial AI technologies [6].

<sup>d</sup>We briefly anticipate and address a potential objection. Namely, "Couldn't these pitfalls also be avoided by simply regulating these for-profit companies?" This is an approach that is already undertaken, so we can provide an evidence-based reply. For-profit companies can and will lobby against this type of regulation in order to minimize, or in some cases, nullify their effects. For an in-depth analysis of the tech lobby's power to limit and prevent regulation of their technology, see Popiel [7].

AI could reasonably be considered to be within the purview of ethical AI. For example, reasoning with ethical principles (e.g., the Doctrine of Double Effect) is certainly part of ethical AI research [8–10].

Our focus in the present paper is on *equitable* AI, which is a subfield of ethical AI. The goal of equitable AI is summarized well in a World Economic Forum white paper on equity and inclusion in AI. They state that “it is critical to re-evaluate the way in which AI is both designed and deployed to ensure that all affected stakeholders and communities reap the benefits of the technology, rather than any harm.” [11] We will expand on this statement; to wit, we consider equitable AI to include two major components.<sup>e</sup> Equitable AI must be:

- *designed*, both by and for, a diverse and inclusive populace, and
- *deployed* such that the benefits are available to all and the harms do not disproportionately affect any particular group.

While one could make a purely moral argument for why AI development should be equitable, there is also a pragmatic justification. That is, much AI research in the United States is publicly funded. It is conducted in academic labs which receive grant funding from federal agencies (e.g. the Department of Defense); money which is collected from the tax payers. Furthermore, private research is able to benefit from this publicly-funded research, since, by law, such research must be made freely available [13]. Since everyone’s taxes are going toward this research, it is only fair that everyone benefits.<sup>f</sup>

One overtly relevant area to equitable AI is the field of AI alignment [14,15]. In broad strokes, the goal of AI alignment is to develop techniques which ensure that the intentions and goals of AI agents are aligned with human values and interests. However, a natural question emerges: *whose* values should such an AI agent be aligned with? There is no universally agreed upon set of ethical principles or norms by which an AI agent could be aligned to, nor is it plausible that such a set will ever be created. However, alignment techniques could be relevant if they were engineered to prioritize equity and fairness.

Finally, we note that equitable AI is certainly related to the notion of “democratized AI” discussed in Montes & Goertzel [2] in the context of decentralized AI architectures.

### 3. Potential Paths Towards Ethical AGI

#### 3.1. Path I: Decentralized AI Architectures

Montes & Goertzel [2] argue that a decentralized approach to AI/AGI development will mitigate the risks posed by centralized, for-profit mega-corporations. They claim that decentralization will enable the masses to compete with mega-corporations, which possess massive amounts of capital and labor which can be directed toward AI/AGI development. They also argue the necessity for AGI “to account for the whole range and gestalt of human ability rather than a minuscule portion of it.” [2] They claim that decentralization enables everyone to participate, thereby ensuring that a diverse population of individuals are able to contribute to AI/AGI development.

However, such an architecture inherently relies on a currency based on distributed ledger technology, such as blockchain. In SingularityNet, the decentralized architecture discussed in Montes & Goertzel [2], a native token is used to buy and sell AI services. They argue that the token “contributes to the robustness and survival of the network by . . . making the network globally open without being tied to any external economies, which could make the

<sup>e</sup>Note that there does not exist a generally accepted definition of equitable AI. For a discussion of suggested definitions, we point the interested reader to Goud et al. [12].

<sup>f</sup>Thank you to an anonymous reviewer for a suggestion which inspired this paragraph.

network vulnerable to manipulation by elites in those unrelated economies.” [2] However, they also indicate that users can buy into and out of the network using major currencies such as USD and EUR. Therefore, while not directly tied to any one economy, the network does appear to be at risk to market manipulation via those external economies which can trade their currency for the native coin. Of course, the alternative — airgapping the network from all external economies, and requiring that currency be earned and spent only within the network — is impractical.

One other important issue is also seemingly not addressed by this approach. That is, while democratizing access is stated as a goal, it isn’t clear how it will be achieved. Montes & Goertzel state that developers can “receive compensation for their work at a market price of their choosing . . . and transact with whom they wish in an open market.” [2] Therefore, those without sufficient capital to purchase AI services at the market rate would be required to sell their labor to earn tokens. Their labor, presumably, would also be compensated at a market rate. The problem is, while not profit-driven, it is not clear how the decentralized approach will ensure that a fair wage is paid for labor, without some centralized actor imposing regulations (e.g., minimum wage).

Finally, we note that there are projects which focus primarily on decentralized *development* of AI and others on decentralized *deployment* of AI. This section has largely considered projects such as SingularityNet [2] which focuses on decentralized deployment. However, projects that focus solely on decentralized development through open source governance (with oversight and cooperation from community collaborators) may also fit within this pathway for equitable AI. It should be noted that a negative consequence of only decentralized development — but not deployment — is that while the implementation may be produced in a democratic manner, access to the end product may be constrained by the end-user’s computing and financial resources.

### 3.2. *Path II: The Capped-Profit Business Model*

A capped-profit business is essentially identical to a for-profit business aside from one characteristic: the amount of profit that any particular investor can receive is capped, usually as a percentage of return on investment. One well-known AI company has pursued this business model: OpenAI. The company was founded as a nonprofit organization [16], but recently transitioned to a capped-profit business model [17]. Specifically, they set a cap of a 100x return on investment. This raises the first issue: the investment cap. If it’s set too low, investors won’t be motivated to invest. Too high, and it’s essentially a for-profit business. This is arguably the case for OpenAI, although they claim that “if [they] are successful, [they] expect to generate orders of magnitude more value than [they’d] owe to people who invest.” [17] Regardless, until that day arrives (if it ever does), the cap does fairly little to reduce the risks that traditional for-profit businesses pose to equitable AI/AGI development.

In their statement announcing their transition from nonprofit organization to capped-profit business, representatives for OpenAI said:

We want to increase our ability to raise capital while still serving our mission, and no pre-existing legal structure we know of strikes the right balance. [17]

During their time as a nonprofit organization, OpenAI was reliant primarily on venture capital and charitable contributions. As they state, they reached a point where they believed they couldn’t raise enough capital without bringing on investors.

. . . we also realized our original structure wasn’t going to work—we simply wouldn’t be able to raise enough money to accomplish our mission as a nonprofit. [18]

However, this statement implies a false dichotomy. Namely, that the choice is between a nonprofit organization — reliant on charitable contributions — or a for-profit business — reliant on (and beholden to) investors. There is another option which can enable the generation of sufficient revenue without reliance on outside investment: the not-for-profit business model.

### **3.3. *Path III: The Not-For-Profit Business Model***

We first note that not-for-profit businesses are distinct from traditional nonprofit organizations in several significant ways. Chiefly, as Hinton states, “they are mostly or totally financially self-sufficient through the sale of goods and services, rather than depending on charitable contributions.” (pg. 14, [19]) Therefore they are able to more easily hire and retain wage-earning employees than nonprofit organizations, which are generally more reliant on volunteer labor.

The Mozilla Corporation is an example of a business which employs this model. In the announcement of its creation, representatives stated:

Any profits made by the Mozilla Corporation will be invested back into the Mozilla project. There will be no shareholders, no stock options will be issued and no dividends will be paid. The Mozilla Corporation will not be floating on the stock market and it will be impossible for any company to take over or buy a stake in the subsidiary. [20]

OpenAI asserts that protections in their charter, which investors must agree to, will protect their development of AGI from for-profit interests. However, in our opinion, the existence of for-profit interests in the development of AGI poses an unacceptable risk, especially if, as OpenAI affirms [18], AGI poses an existential risk to humanity.

The not-for-profit business model fundamentally eliminates this risk. It would enable an AGI company to reliably sustain themselves while being able to assuredly stay true to their goal of equitable AGI. Of course, it isn’t a magical panacea. Such a company would need to produce software and/or services, during AI and (eventually) AGI development, that it could sell to generate revenue to support its research endeavors. However, the benefits of such a structure may very well be worth the cost, especially when the stakes are so high.

One final benefit is that, unlike in the decentralized approach (and like the for-profit model), having a revenue stream would enable an AGI company to offer some services for free. This could be supported by charging for advanced services or offering all services for free up to some usage limit, charging for usage beyond the limit. This would truly democratize access to AI/AGI, enabling even those without sufficient capital to experiment with and contribute to AI models.

### **3.4. *Path IV: An Independent Agency or Government-Run Program***

The rapid development of AI has been compared to a “space race” [21,22], “cold war” [23], and even an “arms race” [24–26] between great powers in both news and academic literature, particularly in international relations and AI ethics spheres. Indeed, some countries have already enacted legislation reflecting this position [27,28]. This characterization has received some direct criticism on ethical grounds that it promotes short-term economic interests over the actual needs of humanity at large [22]. Additionally, Bryson and Malikova [23] provide an analysis of public data on AI investment by country which concludes the narrative is currently overblown.

It is not, however, unthinkable that rising global tensions may bring about a scenario closer to an international AI race, with the end goal of AGI and beyond, particularly when

looking at policy of governments involved in setting “AI supremacy” as a priority [27,28]. A potential catalyst for the exacerbation of this scenario is a rise of tensions leading to fear of a winner-takes-all scenario [24], in which the first country to reach AGI uses this technology to shut out competitors before they develop their own either with force or market dominance.

In the event of such a situation the creation of an independent government agency or state-sponsored entity for the development of AI and AGI would be likely, as from a winner-takes-all scenario, being surpassed becomes an existential risk. This is quite similar to the creation of highly centralized state-sponsored space agencies (e.g., NASA, Roscosmos, etc.) during the space race. Even without the advent of such a scenario that demands the creation of these entities, there are numerous benefits to this government-funded top-down approach, and does not have to presuppose the existence of adversarial relationships between states to kick-start these entities.<sup>g</sup>

The foremost benefit of this approach is the potential for budgetary freedom due to the lack of a profit motive. This would enable funding to flow to the research with the most long-term potential, as opposed to whichever subareas of AI are most easily monetized at the time. Furthermore, the centralization of a state’s AI research into a single organization offers additional benefits. First, the ability for the organization to act as a focal point for setting large scale research priorities. Additionally, such an organization could serve as a regulatory body for AI as well as a center for organizing grants and partnerships. By centralizing these enterprises, the diversity of participants and beneficiaries could be more easily evaluated than under current frameworks such as the far less centralized U.S. National AI Initiative, outlined in H.R.6216 [28].

We give three potential drawbacks of this top-down government-lead centralized approach. First, given the bureaucratic nature of large organizations [29], the ability for a single entity to unilaterally set large-scale research priorities can be detrimental if not handled prudently. The involvement of government funding via tax dollars makes the organization inherently political, which, in combination with bureaucracy, can lead to misplaced priorities where political appearances and favors are prioritized above research goals. In terms of AI safety, this approach is as good as the government behind the organization and the polity behind the government. Ideally it places the development and regulation in the hands of elected representatives who represent the polity who would have the interests of humanity at large; at worst the government has no interest in the polity or humanity at large.

This leads naturally to our second drawback: governments, like for-profit companies, may have purely self-serving interests which conflict with those of some of the polity or other people internationally. However, it’s important to note a distinction: democratic governments at least allow their constituents to vote, protest, etc. in order to allow their opinions to be heard. These rights are not guaranteed when it comes to private companies; in fact, typically people outside of a company have virtually no power whatsoever to influence its decisions (other than through methods provided by their government, e.g. protesting).<sup>h</sup>

Finally, in the event that fear of a winner-takes-all scenario leads to an AI race, the development of AI may not keep pace with the development of AI safety, leading to a dangerous outcome in which the interests of humanity at large are overlooked to great detriment.

### 3.5. *Path V: Academic Laboratories*

The final path we discuss herein is the most traditional means through which research has primarily been conducted: academic laboratories. They offer many benefits in terms

---

<sup>g</sup>Although it is likely that truly large funding would be contingent on such adversarial relationships.

<sup>h</sup>Thank you to an anonymous reviewer for a suggestion which inspired this paragraph.

of ensuring that AI/AGI research is safe and equitable. First, research can be conducted without external funding. While academic research is commonly supported by grants, it is not necessary. Professors can support themselves by fulfilling their teaching and service obligations to their university. Likewise, graduate researchers can be supported on teaching assistantships. In this setting, there is minimal risk that for-profit interests will impact the direction of the research.

Furthermore, tenure has long served to protect the academic freedom of professors. It enables academic researchers to explore topics which may be controversial or otherwise at odds with orthodox thinking [30]. While tenure has its share of critics [31], others feel it can and should be reformed in ways that would improve its ability to produce equitable outcomes. Skoble suggests that “hiring committees as well as tenure committees need to be as sensitive to ideological diversity (and socioeconomic class diversity) as they have become to other dimensions of diversity.” [30]

One challenge to AI research — and especially AGI research — in academia is the difficulty of deep collaboration between labs. In particular, the field of Cognitive Science aims to conduct a multidisciplinary study of cognition via interplay of its six constituent fields of study.<sup>i</sup> While some argue that the field is thriving [34], others have identified challenges of establishing a cohesive research paradigm in academia [35].

More broadly, other studies have identified several challenges to deep interdisciplinary collaboration in academic settings [36,37]. In a survey of faculty at Pennsylvania State University, 96.2% of respondents affirmed that “people issues” were a major hindrance to effective collaboration [36]. The same study concluded that attempts at interdisciplinary research “often [consisted] of faculty continuing piece-meal contributions independent of one another” [36], as opposed to deep collaborations. This is a major concern for effective AGI research, which will undoubtedly require a profound synergy of several disparate strands of research within Cognitive Science, and potentially other fields. This concern is less prevalent in some of the other paths we’ve considered heretofore; e.g., in industrial settings, managerial hierarchies exist which serve (ideally) to construct teams which can collaborate effectively.

#### 4. Classifications of Existing AI/AGI Projects Into Our Framework

Multiple projects working toward the development of more advanced AI are able to be classified within our framework for equitable AI pathways. We provide concrete examples by classifying existing AI/AGI projects into one of our equitable AI pathways, and discuss the clusters.

##### 4.1. Data

Fitzgerald et al. [38] created a dataset of 72 non-narrow AI projects<sup>j</sup> and classify them with respect to seven attributes including (1) the type of institution in which the project is based and (2) the extent of the project’s engagement with AGI safety issues. The notion of *project* here includes companies or organizations which have expressly stated AGI as a goal, and standalone initiatives such as those created by academic labs or governments. The dataset for this work is updated to remove projects that have become defunct or changed direction since the publication of Fitzgerald et al. [38], and updated in the case of projects that have re-branded or changed affiliation. In total, the new cleaned dataset contains 42

<sup>i</sup>The fields contained in the original “Cognitive Science Hexagon” [32] are, alphabetically: Anthropology, Computer Science (some modern versions of the Hexagon replace this with “Artificial Intelligence” e.g., [33]), Linguistics, Neuroscience, Philosophy, and Psychology.

<sup>j</sup>Their exact working definition of AGI is “[AI] that can reason across a wide range of domains.”

projects of the 72 from Fitzgerald et al. [38]. Table 1 lists all of the projects, classified using the criteria given in §4.2. More information on the projects is available at <https://github.com/RAIRLab/Equitable-AI-Development-Paths>.

**Table 1:** The distribution of current AGI development projects with respect to our proposed equitable development pathways.

Decentralized	Capped-Profit	Non-Profit	Government	Academic	For-Profit	Unknown
ANSNA	OpenAI	Whole Brain Architecture Initiative	China Brain Project	ACT-R	AGI Laboratory	Big Mother
Binary Neurons Network			Research Center for Brain-Inspired Intelligence	AERA	AIBrain	Mondad
Brain Simulator II				AIXI	Aigo (Formerly AGi3)	Optimizing Mind
Drayker				Blue Brain	Cyc	Susaro
Human Brain Project				CLARION	DeepBrainz	
MARAGI				FLOWERS	DeepMind	
NDEYSS				LIDA	GoodAI	
OpenCog				NARS	Graphen	
SingularityNet				SOAR	Intelligent Artifacts	
					MSRAI	
					Mind Simulation	
					Mindtrace	
					NNAISENSE	
					New Sapience	
					Olbrain	
					Sanctuary AI	
<b>Count: 9</b>	<b>Count: 1</b>	<b>Count: 1</b>	<b>Count: 2</b>	<b>Count: 9</b>	<b>Count: 16</b>	<b>Count: 4</b>

#### 4.2. Classification Criteria

We classify each project as one of seven categories. The first five are relaxations of our five paths above. By *relaxation*, we mean that projects need not meet all of the lofty criteria we described. For example, Category 3 includes both traditional nonprofit organizations and not-for-profit businesses, while Path 3 focused on the latter. The last two categories are for-profit and unknown. We outline the broad criteria for these classifications as follows.

- (1) Path 1 - Decentralized: The project is either run on a decentralized network (such as blockchain) or is fully open source. In the case that the project is an organization that open sources parts of their code base but not all of it, we do not count them as being truly decentralized.
- (2) Path 2 - Capped-Profit: The project is a “capped-profit” company which caps the maximum profit available to investors.
- (3) Path 3 - Non-Profit: The project is either a nonprofit organization or a not-for-profit company, where all income is reinvested into the organization.
- (4) Path 4 - Government: The project is either an organization run by a government or is an individual project headed fully by a government agency. Note that we do not count government grants for projects under this classification.
- (5) Path 5 - Academic: The project is either an organization run by an academic institution, or an independent project being worked on by an academic lab.
- (6) For-Profit: The project is explicitly designed inside of a for-profit organization to

- be sold, or is a standalone project with the end goal of making profit.
- (7) Unknown: The project does not release enough information to classify it, or it falls outside of all other categories.

### 4.3. Discussion

We note that most of the projects in the decentralized pathway are decentralized development rather than decentralized deployment, relying on open-source community governance. The majority of projects that would be considered equitable under our classification fall into our academic laboratories pathway. However the largest category is for-profit. Therefore, if one agrees with our assertion that for-profit interests pose a risk to AGI development, then this data shows empirically that the discussion herein of alternative pathways is both necessary and timely. Finally, we note that the only two government projects in our dataset both belong to China.

## 5. Conclusion

We considered several development paths for AI/AGI, with an eye to each path’s ability to achieve an equitable distribution of benefits of the technology. None of the paths discussed herein are perfect; each of them has benefits which come at a cost. The best path for equitable AI/AGI could be one still yet to be established, although we believe it is likely to be one of those discussed herein, or possibly a combination of paths, e.g., a largely decentralized architecture which has minimal regulatory oversight by a not-for-profit business, in collaboration with independent government agencies, academic laboratories, etc.<sup>k</sup> While neither the best path, nor the most likely to eventuate are clear, we believe it is crucial that the path taken is one not motivated by profit. Therefore further discussion and implementation of non-profit-driven AGI ventures is paramount to the creation of equitable AGI.

## References

1. OpenAI, *GPT-4 Technical Report*, tech. rep. (2023), <https://arxiv.org/abs/2303.08774>.
2. G. A. Montes and B. Goertzel, Distributed, Decentralized, and Democratized Artificial Intelligence, *Technological Forecasting and Social Change* **141**, 354 (2019) .
3. E. M. Bender, T. Gebru, A. McMillan-Major and S. Shmitchell, On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?, in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, (ACM, New York, NY, 2021).
4. K. Martin, Google Research: Who Is Responsible for Ethics of AI?, in *Ethics of Data and Analytics*, (Auerbach Publications, 2022) pp. 434–446 .
5. T. Simonite, What Really Happened When Google Ousted Timnit Gebru, *Wired* (June 2021), <https://www.wired.com/story/google-timnit-gebru-ai-what-really-happened/> .
6. R. Ochigame, The Invention of ‘Ethical AI’: How Big Tech Manipulates Academia to Avoid Regulation, *Economies of Virtue: The Circulation of ‘Ethics’ in AI* (2022) .
7. P. Popiel, The Tech Lobby: Tracing the Contours of New Media Elite Lobbying Power, *Communication Culture & Critique* **11**, 566 (2018) .
8. M. Giancola, S. Bringsjord, N. S. Govindarajulu and C. Varela, Making Maximally Ethical Decisions via Cognitive Likelihood and Formal Planning, in *Towards Trustworthy Artificial Intelligent Systems*, eds. M. I. A. Ferreira and M. O. Tokhi (Springer International Publishing, Cham, Switzerland, 2022) pp. 127–142 .
9. M. Giancola, S. Bringsjord and N. S. Govindarajulu, A Solution to an Ethical Super Dilemma via a Relaxation of the Doctrine of Triple Effect, in *Life-world for Artificial and Natural Systems, Proceedings of the Sixth International Conference on Robot Ethics and Standards (ICRES*

<sup>k</sup>This notion of collaboration across paths has received support recently; see Marcus & Reuel [39] for a discussion with a particular focus on international cooperation.



- 2021), eds. S. Bringsjord, M. Tokhi, M. Ferreira, N. Govindarajulu and M. Silva (CLAWAR, London, UK, July 2021).
10. M. Giancola, S. Bringsjord, N. S. Govindarajulu and C. Varela, Ethical Reasoning for Autonomous Agents Under Uncertainty, in *Smart Living and Quality Health with Robots, Proceedings of the Fifth International Conference on Robot Ethics and Standards (ICRES 2020)*, eds. M. Tokhi, M. Ferreira, N. Govindarajulu, M. Silva, E. Kadar, J. Wang and A. Kaur (CLAWAR, London, UK, September 2020).
  11. Global Future Council on Artificial Intelligence for Humanity, *A Blueprint for Equity and Inclusion in Artificial Intelligence*, white paper, World Economic Forum (June 2022), [https://www3.weforum.org/docs/WEF\\_A\\_Blueprint\\_for\\_Equity\\_and\\_Inclusion\\_in\\_Artificial\\_Intelligence\\_2022.pdf](https://www3.weforum.org/docs/WEF_A_Blueprint_for_Equity_and_Inclusion_in_Artificial_Intelligence_2022.pdf).
  12. S. Goud, A. F. Mertz, J. L. Sheats and D. Chou, *A Blueprint for Equitable AI: Building and Distributing Artificial Intelligence for Equitable Outcomes*, tech. rep., Aspen Institute Science & Society Program (2023), <https://www.aspeninstitute.org/publications/blueprint-for-equitable-ai/>.
  13. A. Nelson, Ensuring Free, Immediate, and Equitable Access to Federally Funded Research, *Office of Science and Technology Policy* (August 2022), <https://www.whitehouse.gov/wp-content/uploads/2022/08/08-2022-OSTP-Public-Access-Memo.pdf>.
  14. N. Soares and B. Fallenstein, Agent Foundations for Aligning Machine Intelligence with Human Interests: A Technical Research Agenda, *The Technological Singularity: Managing the Journey*, 103 (2017).
  15. N. Soares, The Value Learning Problem, in *Artificial Intelligence Safety and Security*, (Chapman & Hall/CRC, 2018) pp. 89–97.
  16. G. Brockman and I. Sutskever, Introducing OpenAI, *OpenAI Blog* (December 2015), <https://openai.com/blog/introducing-openai/>.
  17. G. Brockman and I. Sutskever, OpenAI LP, *OpenAI Blog* (March 2019), <https://openai.com/blog/openai-lp/>.
  18. S. Altman, Planning for AGI and Beyond, *OpenAI Blog* (February 2023), <https://openai.com/blog/planning-for-agi-and-beyond/>.
  19. J. B. Hinton, Relationship-to-Profit: A Theory of Business, Markets, and Profit For Social Ecological Economics, PhD thesis, Stockholm Resilience Centre, Stockholm University, (Stockholm, Sweden, 2021).
  20. MozillaZine, Mozilla Foundation Announces Creation of Mozilla Corporation, *MozillaZine* (2005), <https://web.archive.org/web/20060907025204/http://www.mozillazine.org/talkback.html?article=7085> (Last Accessed 2/21/23).
  21. J. R. Allen and A. Husain, The Next Space Race is Artificial Intelligence, *Foreign Policy* (November 2017).
  22. I. Ulnicane, Against the New Space Race: Global AI Competition and Cooperation for People, *AI & SOCIETY* (2022) <https://doi.org/10.1007/s00146-022-01423-0>.
  23. J. J. Bryson and H. Malikova, Is There an AI Cold War?, *Global Perspectives* **2**(06 2021), 24803 <https://doi.org/10.1525/gp.2021.24803>.
  24. N. D. Wim Naudé, The Race for an Artificial General Intelligence: Implications for Public Policy, *AI & SOCIETY* (2019).
  25. M. Auslin, Can the Pentagon Win the AI Arms Race?, *Foreign Affairs* (2018).
  26. J. D. Remco Zwetsloot, Helen Toner, Beyond the AI Arms Race, *Foreign Affairs* (2018).
  27. New Generation AI Plan (2017), [http://www.gov.cn/zhengce/content/2017-07/20/content\\_5211996.htm](http://www.gov.cn/zhengce/content/2017-07/20/content_5211996.htm).
  28. H.R.6216 — 116th Congress (2019-2020): National AI Initiative Act of 2020 (2020).
  29. M. Cote, Why are Government Bureaucracies Inefficient? A Prospective Approach, *SSRN Electronic Journal* (2018).
  30. A. J. Skoble, Tenure: The Good Outweighs the Bad: A Surreponse to James E. Bruce, *Journal of Markets & Morality* **22** (2019).
  31. J. E. Bruce, Does Academic Tenure Promote the Common Good?, *Journal of Markets & Morality* **22** (2019).
  32. G. A. Miller, The Cognitive Revolution: A Historical Perspective, *Trends in Cognitive Sciences* **7**, 141 (2003).
  33. A. Schierwagen, The Way We Get Bio-Inspired: A Critical Analysis, *From Brains to Systems* (2010).
  34. M. McShane, S. Bringsjord, J. Hendler, S. Nirenburg and R. Sun, A Response to Núñez et al.’s

- (2019) “What Happened to Cognitive Science?”, *Topics in Cognitive Science* **11**, 914 (2019) .
35. R. Núñez, M. Allen, R. Gao, C. Miller Rigoli, J. Relaford-Doyle and A. Semenuks, What Happened to Cognitive Science?, *Nature Human Behaviour* **3**, 782 (2019) .
  36. H. Lin, Opportunities and Challenges for Interdisciplinary Research and Education, *Journal of Natural Resources and Life Sciences Education* **37** (2008) .
  37. S. Glied, S. Bakken, A. Formicola, K. Gebbie and E. L. Larson, Institutional Challenges of Interdisciplinary Research Centers, *Journal of Research Administration* **38** (2007) .
  38. M. Fitzgerald, A. Boddy and S. D. Baum, *2020 Survey of Artificial General Intelligence Projects for Ethics, Risk, and Policy*, tech. rep. (2020).
  39. G. Marcus and A. Reuel, The World Needs an International Agency for Artificial Intelligence, Say Two AI Experts, *The Economist* (April 2023), <https://www.economist.com/by-invitation/2023/04/18/the-world-needs-an-international-agency-for-artificial-intelligence-say-two-ai-experts> .

## **ROBOETHICS IN THE LOOP: THE ELS ISSUES IN THE REXASI-PRO PROJECT**

GIANMARCO VERUGGIO and MAURIZIO MONGELLI

*CNR-Istituto di Elettronica e di Ingegneria dell'Informazione e delle Telecomunicazioni  
Corso Perrone 24, 16152 Genova, Italy*

*E-mail: [gianmarco@veruggio.it](mailto:gianmarco@veruggio.it), [maurizio.mongelli@ieiit.cnr.it](mailto:maurizio.mongelli@ieiit.cnr.it)  
<https://www.ieiit.cnr.it/>*

IORELLA OPERTO

*Scuola di Robotica, via Balbi 1A, 16126 Genova, Italy*

*E-mail: [operto@scuoladirobotica.it](mailto:operto@scuoladirobotica.it)  
<https://www.scuoladirobotica.it/>*

In this paper authors describe the methodology for Roboethical analysis and evaluation in the REXASI-PRO project where an Ethics by Design methodology is adopted from the proposal, through the development and testing phase of the system. The analysis of the methodology by which ethical requirements are applied by the REXASI-PRO project, presented here, is valuable in several respects because ELS (Ethical, legal, and Societal) issues have been analysed from the proposal and are discussed with partners at every stage of project development. This methodology seemed to the Partnership to be more comprehensive and adherent to the spirit and letter of recent European ethical recommendations than an approach involving the assessment of ELS requirements intervening in an ex post analysis, and thus influencing the initial and developmental design in a more limited way. For this, a number of information and opinion gathering tools (Structured Interviews, Focus Groups, Surveys, calls for discussions on technical aspects) to be administered to and with designers have been envisaged, with the objectives of verifying the adherence of the project to the ethical requirements and identifying critical points of development.

### **1. Introduction**

We outline in this paper a roboethics analysis applied to a robotics project where an Ethics by Design methodology is adopted since the drafting of the initial concepts. The authors are Partners in a European robotics project that has social and ethical implications, the --PRO (HORIZON-CL4-HUMAN-01-01, RIA- Research and Innovation Action, Project No. 101070028).

The project started on 1.10.2022 with the duration of 36 months and it aims to release a novel engineering framework to develop “ethically confirmed, greener and trustworthy Artificial Intelligence solutions”. The application system is a Multi-Robot Systems and Data Collection Platform supporting people with mobile disabilities gap. Two Case studies are being developed, to confirm the feasibility of the objectives, led by an AI-based orchestrator layer managing the fleet of AI-based swarms: 1. A Smart, Safe and reliable Wheelchair operating in real-life scenario populated by humans; 2. Flying robots, in real-life scenarios, capable of flying autonomously and replacing the need for human intervention within indoor/underground environments. This technology is promising to facilitate the evacuation of people with permanent or reduced mobility in emergency situations, such as a fire in an infrastructure with the risk of collapse.

The process of analysing and evaluating with the Partnership the technical solutions for ethical issues of every phase of the development causes ELS issues like safety, security, issues

of privacy, of user's comfort and well-being and AI explainability to be addressed carefully and during the development itself. [1]

To this end, the REXASI-PRO project introduces several novelties. The project will develop in parallel the design of novel trustworthy-by-construction solutions for social navigations and a methodology to certify the robustness of AI-based autonomous vehicles for people with reduced mobility.

The trustworthy-by-construction social navigation algorithms will exploit mathematical models of social robots. The robots will be trained by using both implicit and explicit communication. REXASI-PRO methodology augments existing system-level and item-level engineering frameworks by leveraging novel explainability methods to improve the entire system's robustness. REXASI-PRO will release additional verification and validation approaches for safety and security with the AI in the loop. Among the other developments, a novel learning paradigm embeds safety requirements in Deep Neural Network for planning algorithms, runtime monitoring based on conformal prediction regions, trustable sensing, and secure communication. The methodology will be used to certify the robustness of both autonomous wheelchairs and flying robots. The flying robots will be equipped with unbiased machine learning solutions for people detection that will be reliable also in an emergency. Thus, REXASI-PRO will make the AI solutions greener. To this end, both an AI-based orchestrator to augment the intelligence of the robots and topological methods will be developed. The REXASI-PRO framework will be demonstrated by enabling the collaboration among autonomous wheelchairs and flying robots to help people with reduced mobility.

## **2. The REXASI-PRO Framework Application**

### **2.1. *Ethics by Design***

Ethical Guidance concerns and are to be applied to all research projects involving the development or/and use of robotics and artificial intelligence (AI)-based systems or techniques.

The reference documents are the "Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment", the Roboethics Roadmap [2] [3] [4] and the DGPR.

The ethical principles that must be protected in REXASI-PRO are:

- respect for human agency (autonomy, dignity and freedom);
- privacy, personal data protection and data governance;
- fairness (Avoidance of algorithmic bias: Universal accessibility; Fair impacts)
- individual, social, and environmental well-being;
- transparency;
- accountability and oversight.

In REXASI-PRO, the aim of Ethics by Design is to embody the ethical principles into each phase of the development process allowing that ethical issues arising in itinere are addressed as early as possible and followed up closely during research activities. [Fig. 1]

In fact, the approach evoked above, with an initial analysis of ethical challenges - which is sometimes necessarily generic - and a final assessment of the implementation of ethical recommendations, may not intervene in the poignant developmental phases and may not involve all partners into the directly participation of the "materialization" of the real requirements in the engineering datum.

The aim of REXASI-PRO Ethics by Design methodology is to make all designers aware of the potential ethical concerns in every phase of their development of the system. The adopted methodology also applies to piloting, to the volunteer's selection, and to assessment and evaluation parameters.

This approach, tailored to the objectives of the research proposed, takes into consideration that, for instance, ethics risks can be different during the research phase from that of the deployment or implementation phase.

Given the Use Cases selected - a Smart Wheelchairs for Safe and Reliable Operations and Flying robots capable of flying autonomously and replacing the need for human intervention within indoor/underground environments - the ethical requirements recommended for the global objectives - and thus standardized for many other cases - are to be assessed considering the end user, the social environment, the physical environment and, last but not, the autonomous machines and the AI programs.

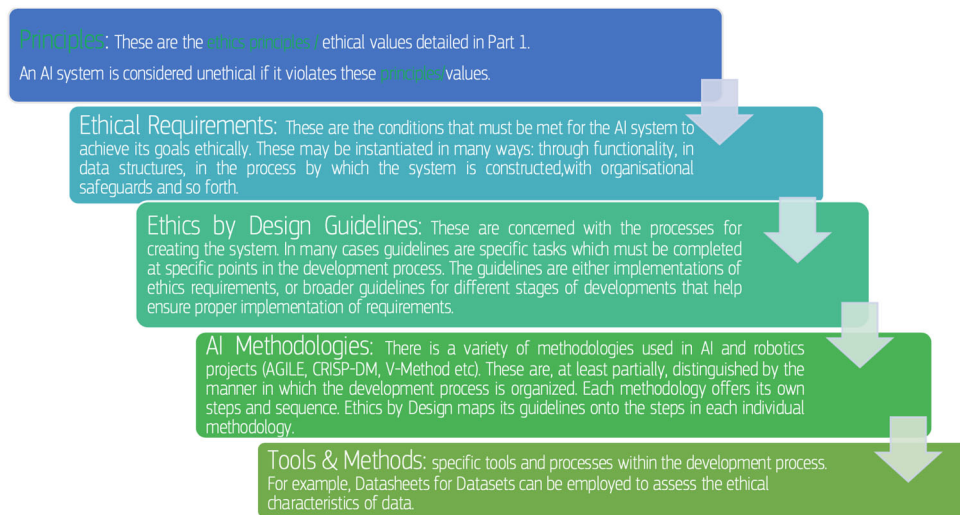


Figure 1: The 5-layer Model of Ethics by Design. The Development Process

### 3. The REXASI-PRO ELS Methodology

In REXASI-PRO, the elements to be analysed to define ethical requirements are:

- the end user
- the health care personnel or persons accompanying the end user
- the environment
- the Multi-Robot System and Data Collection Platform.

The analysis of these elements defines the fine level of ethical requirements.

“People with mobile disabilities gap” identified as the end user(s) could range from a sportsman awaiting meniscus surgery to a person with severe lower limb paralysis and psychological imbalance caused by the pathology. To be adoptable to a large number of cases, REXASI-PRO ethical requirements assessment must consider every potential application, and therefore every potential end user. In the two selected Use Studies, the actual conditions of the user testing the system define the accuracy of ethical requirements.

Similarly, for the staff next to the end user. They may be health care personnel, physicians, and even family members of other patients-if the setting is a hospital or medical centre. The behaviour of people around the system with the end user can range from a level of knowledge of the presence of an autonomous system to misunderstanding what is going on (they see a wheelchair moving autonomously but may think it is controlled by the end user) to recognizing that it is a robot and stopping to look, or asking for information, thus obstructing the path area.

Service robots used in real, human-inhabited and unpredictable environments assisting disabled humans in hospitals, functional recovery centres, or rehabilitation institutions must

meet higher levels of requirements standards compared to other service robots working in less challenging habitats.

In these cases, to the classic parameters of dependability (safety, security, affordability, and maintainability) of each sound engineering project, essential ethical elements are added for each parameter, involving aspects of human-robot interaction, privacy concerns, usability difficulties, and cost-benefit balancing. It must also be ensured that AI programs are free from bias and major errors from inappropriate patterns.

The REXASI-PRO System integrates the requirements for ethical dependability with human-oriented acceptability elements in a real user-centred design, and along all stages. [5]

#### **4. The System and Process Analysis**

##### **4.1. *User-related ethical requirements***

The question the REXASI-PRO Partnership had posed in drafting the proposal is: How the design and development of system could provide value to the user?

Under no circumstances, can the end-user receive pressure to use it; therefore, the system must exhibit features able to improve the user's quality of life in several aspects. Thus, the user is willing to adopt it. Otherwise, the system cannot be considered dependable.

The benefits to the individual and society have been expressed by the field literature and, taking into account the reality of the economic and social situation of the health care sector, European Wide, autonomous machines or autonomous transportation systems can relieve caregivers of various tasks, offering them the opportunity to have more time for tasks that only they could do with patients.

For the individual end user, the use of a robotic system such as REXASI-PRO offers the possibility of greater autonomy and elimination of downtime in waiting for personal care givers when a transfer from one place to another is needed.

##### **4.2. *The end user' psycho-physical comfort***

The system should be designed to ensure the psycho-physical well-being of the end user because ethical requirements demand that he/she feels good and be in a comfortable situation in wheelchair movement.

A person who needs a wheelchair for his or her own movement, whether for a short time or with a permanent disability, whether an athlete who has to undergo ligament surgery or a paralyzed elderly person, experiences an anxious state of mind and will have introjected a general uncertainty about losing control of his or her body and, if this has been happening for some time, will have lost in some way his or her sense of balance and space, and will have interceded in sarcopenia.

The wheelchair should maintain a smooth motion so as not to create stress and a sense of anxiety by avoiding subjecting the end user to jerking, abrupt linear or angular braking or acceleration, or following trajectories that generate fear of falling, lurching, or bumping into walls or people.

There are other anxiety states that may intervene: sense of claustrophobia, if the end user employs seat belts, and dizziness. All of these requirements are not just about the needed safety and security standards, but about the overall sense of comfort that the system should convey to the end user.

In fact, from the point of view of the end user's psychological state, the end user will have to get used to the state of being transported by a machine and not by a human, with whom he or she might chat and with whom he or she might have even a brief but interpersonal exchange relationship. We know that these moments of small talk are often important for patients and manage to relieve a little of the anxiety of the transit that may be, for example, to a problematic analysis in another hospital ward.

This means that the motion planning function not only needs to optimize the engineering parameters of the trajectory, so it can use a trajectory that is unfavourable from the point of view of time and energy consumption, but safeguards the comfort of the passenger.

Indeed, we must take into account that although the end user has been informed of many elements of the overall system behaviour, he or she may not always be confident and may have moments of anxiety and uncertainty about the trajectory.

For example, in view of an obstacle, the system will anticipate the manoeuvre to avoid abrupt movements and at the same time communicate to the passenger that they are aware of the situation and have made the right decisions.

#### **4.3. *The human-robot communication channel***

The element of communication between the machine and the user is crucial, so that the user always has knowledge of what is going on, so that the user feels safe and cared for. [6]

The human-robot communication support can be achieved through a graphic screen showing the route, destination, and even different trajectories. In fact, that the end user knows the subtasks of the mission will have a calming effect: for example, knowing that you will be going to an elevator and going down three floors. The robot will also inform the end user of any path changes. For example, if in the past travels to the same destination, the robot will have chosen one route, and in one case will choose another, which the end user is not used to, the robot will inform the end user of the reason for the change (washing the floor, etc.).

These are behaviours of the robot that make the user much more serene and in a way can also make the transfer enjoyable, which should feel more like a fun and rewarding time as well (after all, it is not every day to be transported by such a sophisticated and intelligent system; somewhat à la Patch Adams).

#### **4.4. *Health professionals and staff, and/or family members***

The benefit to health care workers is that they will be assisted by the system in routine tasks, thus being able to devote themselves to personal care tasks that would be irreplaceable.

The ethical dependability issues here are similar to those for the end user: the caregiver must be informed by the system about autonomous missions, so there must be full task transparency. At the time when the caregiver will "hand over" the patient to the system, there will have to be a handover of information.

The second aspect here is the monitoring of the path from source to destination. The system will need to be able to communicate mission information to a control room and be able to send alerts if for some reason there are major changes in the path, in the behavior of the end user, in the environment itself.

The third is the responsibility ascription problem: who is the human ultimately responsible for the mission? The last caregiver who came in contact with the patient? the control centre?

The control unit could be unified or distributed in a series of checkpoints along the route: when the wheelchair starts the route, it communicates to the control centre where an operator can check the status of the mission. The ethical aspect here also relates to the stress status of the operators, and controlling the route of the vehicle is essential to minimize the caregivers' state of worry. In case of alarm, the nearest caregiver, nurse, is informed.

We are here in the domain of mixed-team, human-robot issues, where human operators must be aware of the robot's behaviour in order to maintain a trust in it. This implies adequate training of all caregivers involved.

Indeed, the rules of a Trustworthy AI stipulate that human caregivers should not be burdened with ethical and legal responsibilities with respect to the robot's behaviour nor should they be under stress where they should instead be relieved of it by the machine. In our project, technologies are expected to improve, not worsen, the conditions of care givers.

Unfortunately, it happens that with the introduction of AI systems, humans are subjected to work rhythms and pressures that are imposed by the rhythms and performance of machines, especially if the introduction of technologies has obviously led to a decrease in staffing levels.

#### 4.5. The REXASI-PRO system for the humans in the environment

In the environment where the end user and the robotics system are located, there will be other people who can be physicians, nurses, other patients or other patients' families; whether at home or in a centre, other patients and family members can move in the environment. All of these people may not be aware of a robot's behaviour and motion. The system must also take into account and respect these people because their behaviour can interfere with that of the robot and can worry the end user.

In turn, the behaviour of these third parties must take into account the presence and motion of the robot and for this they must be informed with some well-designed messages.

It is necessary to design along the path and from the beginning to the destinations a series of conditions that inform and see an operational code for the resolution of motion conflicts (e.g. the precedence, etc) as they already exist in the hospitals to divide the routes between the exterior visitors and the staff and patients; or which differentiate between stretcher bearers the patients heading towards the operating rooms; or signalling the presence of radioactive material. These cues can be arrows, floor markings, light signals on signs, etc. Where necessary, special routes for the robot and emergency situations should be provided in case of an alarm.

#### 4.6. Issues of ethics of proximity

From the perspective of wheelchair trajectory, we need to consider not only the inherent engineering and robotic aspects of movement, but also important issues of proximity etiquette that work among humans. Although this should be at the expense of the economy of movement, the wheelchair will not pass between two people chatting and will try wherever possible not to get behind other people along the way. [Fig. 2]

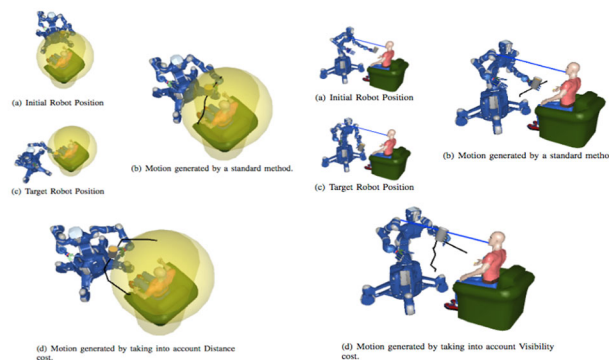


Figure 2: Socially Intelligent Motion Planning of a Humanoids  
Credits: European Project PHRIENDS

Humans perform tasks and move, maintaining an area of safety space around them that ensures a sense of their own autonomy in space. Also in REXASI-PRO; the end user's space (the system end user/wheel chair) should not be crossed. [7]

There will be situations when the robot will advance empty, to approach the patient: in this case we must expect that people in the environment will stop to look at an empty wheelchair walking by itself and may form a small crowd or otherwise get in the way. As well as predict that a child will try to stop or sit down or, if the system will navigate in an environment such as a garden, etc., there may be animals.



#### **4.7. *The robotics divide***

Among the general ethical problems posed by the introduction of the system into society we must anticipate issues of robotics divide, relating to those nations or communities that will not be able to afford to buy and maintain the robot. Issues of generational divide, in the sense that elderly people will be less likely to use the robot.

And issues relating to the loss of jobs, job displacement, for the operators who will be replaced by the robot and who will be, as often happens, the least qualified and least ready to find another job.

### **5. Context Defined Ethical Challenges**

In designing the REXASI-PRO system according to the "Ethics by Design" methodology as in our proposal, several challenges are to be encountered in applying the expected ethical requirements. Our unified approach is to analyse these challenges in collaboration with all the designers involved in different aspects of the work. Therefore, it is important to know the expected level of attention with respect to ethical requirements at the topical moments of the system design.

The in itinere tools for collecting design information and interacting with designers serve primarily to align the latter with the ethical requirements in the selected Use Case. One of the key points concerns the concept of what we can call Ethical Dependability.

Acceptability and dependability are fundamental elements of the robotics development and design engineering process because they influence the reliability and success of the use of a robot in real life conditions. They also influence the acceptance of the robotics system, its correct use and its duration.

Elements of the robot's functional modules and exterior aesthetical design are also important because they can promote or discourage the use of the robot by end users, and by humans directly interested.

The know concept and scope of Dependability (safety, security, affordability, and maintainability) shall involve here elements of ethics and thus declined also on aspects of Human-robot Interaction, Comfort Psychology, attention to Privacy and ethics of Proximity, privacy concerns, usability difficulties, and cost-benefit balancing. Elements that an engineering design of another kind may not consider.

Therefore, circularity of interaction between ethical requirements experts and robotic and AI planners will be important, to refine directions once the results of the Surveys and focus groups have been collected (reflexive loop). [8]

### **6. Conclusions**

In operating a multi robot system for Assistive robotics like REXASI-PRO, the ethical requirements are related to the need of the assisted person with reduced mobility and operating in an unstructured environment inhabited by humans who are not necessarily aware of the action of the robotics system.

We have considered some challenging environmental characteristics for such a system.

If we consider, for example, that the system is to operate in a hospital and is to autonomously transport the user through several floors and various corridors, we must provide a control centre that supervises the path and can alert some control points located along the way.

If we design that the system must operate outdoors where video cameras cannot be placed, or where their reliability is uncertain, the use of the drone could be necessary.

If, in addition, we anticipate that the user has reduced mobility and also emotional-cognitive problems, the challenges become greater.

The project's central approach is "Ethics by Design". [9] This indicates that the project addresses ethical issues from the beginning ensuring that ethical principles guide the development of design and followed up closely during research activities. This approach should

be modulated according to the limits and frame of the project, to the environment and to the challenges.

Moreover, ethical requirements can differ from the research phase and the deployment with respect to the implementation phase. [10]

It is important that ethical principles are incorporated into the design without altering their weight by finding at the same time the best engineering solutions to ensure the most effective functioning of the system.

## 7. References

1. Carlevaro A., Mongelli M. (2022). "A New SVDD Approach to Reliable and Explainable AI". *IEEE Intell. Syst.* 37(2): 55-68.
2. Veruggio G., "The EURON Roboethics Roadmap", (2006) 6th IEEE-RAS International Conference on Humanoid Robots, pp. 612-617.
3. Veruggio G., Abney K. (2011). "Roboethics : the applied ethics for a new science"- In Lin P. and Abney K. (Eds), *Robot Ethics*, MIT Press, Cambridge, p. 355.
4. Veruggio G., Bekey G., Operto, F. (2016). "Roboethics: Social and Ethical Implications", in B. Siciliano e O. Khatib (Eds), *The Springer Handbook of Robotics*, pp. 2135-2160.
5. Operto F. (2011). "Ethics in Advanced Robotics". in *IEEE Robotics & Automation Magazine – Special Issues on Robot Ethics, Robotics and Automation Society*, n. 18, v. 1, anno 2011, pp. 72-78.
6. Speranza S., Recchiuto C.T. , Bruno B., and Sgorbissa A. (2020). "A Model for the Representation of the Extraversion-Introversion Personality Traits in the Communication Style of a Social Robot," 2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN), Naples, Italy, 2020, pp. 75-81.
7. Clodic A., Fleury S. , Alami R., Chatila R., Bailly G., et al.(2006). "Rackham. An Interactive Robot-Guide". *IEEE International Workshop on Robots and Human Interactive Communications (ROMAN)*, Sep 2006, Hatfield, United Kingdom. pp.502-509.
8. Narteni S., Orani V., Vaccari I., Cambiaso E., and Mongelli M. (2022). "Sensitivity of Logic Learning Machine for Reliability in Safety-Critical Systems". *IEEE Intell. Syst.* 37(5): 66-74 (2022)
9. Veruggio G., Operto F., "Roboethics: a bottom-up interdisciplinary discourse in the field of applied ethics in robotics", in "Machine Ethics and Robot Ethics", pp. 79-85, Routledge, 2020.
10. Tamburrini G. (2009). "Robot ethics: a view from the philosophy of science", in R. Capurro, M. Nagenborg (Eds), *Ethics and Robotics*, IOS Press, Heidelberg, 2009, pp. 11-22.

## 8. Acknowledgment

"The work was supported in part by REXASI-PRO H-EU project, call HORIZON-CL4-2021-HUMAN-01-01, Grant agreement no. 101070028"  
<https://rexasi-pro.spindoxlabs.com/>

## 9. Disclaimer

"Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or [name of the granting authority]. Neither the European Union nor the granting authority can be held responsible for them."

## PROSPECTIVE INSIGHTS FROM THE REAR-VIEW MIRROR: REVISITING THE MONARCH PROJECT

MARIA ISABEL ALDINHAS FERREIRA

*Centre of Philosophy of the University of Lisbon. Faculdade de Letras, Alameda da Faculdade. Lisbon.  
Portugal*

and

*Institute for Systems and Robotics, Instituto Superior Técnico. University of Lisbon.  
Portugal. Av. Rovisco Pais 1, 1049-001 Lisbon, Portugal*

*E-mail: [isabelferreira@letras.ulisboa.pt](mailto:isabelferreira@letras.ulisboa.pt)*

At a time, design in its dual functional and aesthetic facets is assumed, more and more, as an imperative in the production of robots meant to interact with human beings, the present paper revisits the process that led to the design and deployment of the social robot produced in the context of the MONARCH project<sup>\*</sup>, nearly a decade ago. Taking what we designated as a rear-view perspective, the paper highlights the essential roles played by the concepts of **user-centred design** and that of **user's framework of reference** in that project. From the evidence provided by that context of experience, we draw insights that may prove to be useful for the present and future practices.

### 1. Introduction

Scientific and technological endeavour continually advances and accumulates over time. The cumulative character of human knowledge is driven by an iterative process that involves formulating hypotheses, conducting experiments, analysing and assessing data, refining theories and redefining practices. These are the main stages of a dynamics that fuels technological advancement and is responsible for its evolution, leading to the development of new systems, new machinery, new processes.

Social robotics has seen a remarkable progress in the last decades, (i) from the days of early robotic systems with very primitive basic behaviors<sup>†</sup> (ii) to robots capable of some interaction according to a small set of defined behavioural and communicative patterns<sup>‡</sup>, (iii) to progressively more autonomous systems, capable of learning from the prompts made by the

---

<sup>\*</sup> European project FP7 ICT-9-2011-601033 (MOnarCH)

<sup>†</sup> Cf on this purpose the work of William Grey Walker. <http://www.rutherfordjournal.org/article020101.html>

<sup>‡</sup> One of the leading figures, responsible for advances in the field of social robotics, is Kerstin Dautenhan. For an extensive listing of some of her most relevant titles on the topic visit: <https://dblp.org/pid/44/4107.html>

surrounding environment and consequently capable of navigating and interacting adequately in domestic or public spaces, as it is the case, for example, of Airstar at Seoul airport<sup>§</sup>.

In Europe, much of this technological progress has been fostered for decades by frameworks as those of the EU 7<sup>th</sup> Program for Research and Innovation- fp7 (2007- 2013), Horizon 2020 (2014-2020) and Horizon Europe (2021-2027) that allowed for the construction, trial and deployment of social robot prototypes and their progressive incorporation in diversified social contexts, as those of care houses, schools, hospitals and public spaces, performing distinct roles and consequently being also somehow functionally diverse.

Though the projects that were developed within these frameworks have generally met their specific goals, progressing according to the planned milestones, we cannot help feeling that a bird's eye view, an overall critical appraisal on how design options have evolved throughout time, identifying those that were discarded during this evolutionary process and those that progressively became more sophisticated, would be very beneficial for all that are involved in their design and production.

Bearing that in mind, and nearly a decade after its conception, production and deployment, we have decided to revisit the fp7 MOnarCH project\*\* bringing forth its main theoretical and methodological pillar- the concept of User-Centred Design and what was at the time defined as the User's Framework of Reference<sup>††</sup>, hoping in so doing, to get useful insights to today's practices.

The next section will briefly approach the main guidelines that give substance to the concept of User- Centred Design.

## **2. User-Centred Design: A Design Methodology**

User-Centred Design (UCD) is a working methodology to product design and development that assumes as essential design guidelines those determined by the needs, goals, and life contexts of the potential users/consumers and conceives and produces artifacts taking this information into consideration. It involves understanding who the target users will be and their framework of reference, Ferreira [1] , [2] , designing accordingly and following an iterative process in which options are tested and updated according to the user's feedback - user's experience (UX). The concept of UCD has, for decades, been applied by multiple domains of industry, namely automobile industry, where customer satisfaction requires a good balance between design quality and price and resides not only in safety standards, but also in the aesthetics and, whenever possible, in a certain degree of customization.

The concept of User- Centred Design was coined during the 1980's and developed in the subsequent decades due to a confluence of factors:

---

<sup>§</sup><https://www.google.com/search?q=robot+seoul+airport&oq=robot+at+seoul+&aqs=chrome.1.69i57j0i22i30.3529769942j0j15&sourceid=chrome&ie=UTF-8#fpstate=ive&v1>

\*\* FP7-ICT-9-2011-601033 (MOnarCH)

†† Cf Ferreira (2015)

1. The Participatory Design Movement that originated in Scandinavia in the 1970s and that emphasized the need to involve end-users in the design process in order to elicit their insights relatively to the product and their preferences.
2. Work at Xerox PARC: Xerox's Palo Alto Research Center (PARC) was also instrumental in advancing user-centered design principles in ICTs. Researchers such as Donald Norman [3] and Terry Winograd [4] conducted studies on human-computer interaction and emphasized the importance of designing systems that align with human capabilities and mental models.
3. The release of the Apple Macintosh computer in 1984 showcased a more user-friendly interface, featuring graphical elements and a mouse. This marked a significant shift towards making technology more accessible and intuitive for users.
4. The International Organization for Standardization (ISO) published the ISO 9241 standard series on ergonomics of human-computer interaction. These standards highlighted the importance of considering user needs, usability, and user-centred design principles in the development of interactive systems. It defined the six essential procedures that typify this approach, [5] ISO 9241-210, (2010):
  - a) The design is based upon an explicit understanding of users, tasks and environments.
  - b) Users are involved throughout design and development.
  - c) The design is driven and refined by user-centred evaluation.
  - d) The process is iterative.
  - e) The design addresses the whole user experience.
  - f) The design team includes multidisciplinary skills and perspectives.

The rich multidisciplinary framework necessarily called to be involved in this process led some authors, e.g., Steen [6], Giacomini [7] to propose the term “Human-Centred Design” as more suitable for covering all aspects of what being a human means and not only those specifically concerned with usability. However, as Ferreira [8] points out, either user-centred design or human-centred design highlight the fact that all artifacts, including technological artifacts, are determined in their function and form by the anatomy and physiology of the user, by their psychology and life experience by their expectations towards technology as well as by the specificities of the context of use.

Consequently, rather than expecting people to just adapt to a new technological artifact, learning how to handle or interact with it, robotic engineering must be capable of thinking and anticipating how the system can be designed to best suit the people and the society who need to use it.

In order to achieve this aim, one needs to identify the potential end-user and design for the variability represented in the population, spanning such attributes as age, size, strength, cognitive ability, prior experience, cultural expectations and goals, optimising this way performance, safety and well-being.

### **3. User- Centred Design in the MOnarCH Project**

#### **3.1. The Project’s Aims and Goals**

The MOnarCH Project aimed to develop a system of networked robots (NRS), capable of creating an amusing and stimulating environment for the children of the paediatric

ward of an oncological hospital, contributing to the improvement of their life context by keeping them, as much as possible, enjoyably active. Though it was anticipated that the robots might and probably would also interact with adults (staff and adult visitors), children were identified as the end-users and all the design process was defined taking in consideration their physical and psychological status, meet their preferences and fill their expectations. The social robot would navigate the paediatric ward interacting with children whose age could range from infants to teenagers up to sixteen.

In order to highlight the centrality of the user in the design process and the main parameters to consider, Ferreira [2] defined what she called the “User’s Framework of Reference (UFR).

Table 1: User’s Framework of Reference in the MOnarCH project

Universe	Nature and Status				Typical Environment/ Context	Interaction Frequency	Expected Roles	Robots Functionalities
Plural:  Simultaneous multiple Users	Age  1-16	Gender:  Not Gender specific	Particularities:  Frail condition	Civilizational/cultural context:  European	Institutional Space  Pediatric Ward  Pediatric Ward’s classroom	Regular-Daily	Edutainment	Navigation  Obstacle Recognition and Avoidance  Basic Communicative Behaviors

The MOnarCH robot (Mbot) was intended to interact with multiple users- male and female children experiencing severe health conditions. The robot would be involved in edutainment activities taking place in the pediatric ward and in the pediatric ward’s classroom. These activities would include essentially (i) interacting with children (ii) playing a game with them, and (iii) whenever possible act as school teaching assistants leading the children to the classroom, showing educational videos.

### 3.2. The Conceptual Phase

The definition of an engaging and stimulating visual image of the Mbot for children was set up as a priority. Before initiating the sketching of the outer shell and definition of interfaces and functionalities available, a survey was conducted in two public schools in the suburban area of Lisbon. The goal of this survey was twofold: (i) to verify the existence in children of a prototypical mental visual representation associated to the concept of [robot]; (ii) to identify the main common semantic features sustaining the concept, i.e., the expected functional roles associated to the concept.

The same questionnaire was applied to three distinct gender- mixed groups: one constituted by children aged 8-9 years old, the other constituted by children aged 10-11 years old and finally another one constituted by 13-16 year-olds. These three groups together defined a universe of 120 students.

The inquiry took place in the normal classroom context. Students answered the same questionnaire in their classroom environments, without any previous warning or motivation. All students were given a period of 45 minutes to accomplish the task. Questionnaires were anonymous just referring the age of the inquired<sup>‡‡</sup>.

The survey comprehended two distinct tasks: 1- Answering a four question inquiry 2- drawing what their robot would look like.

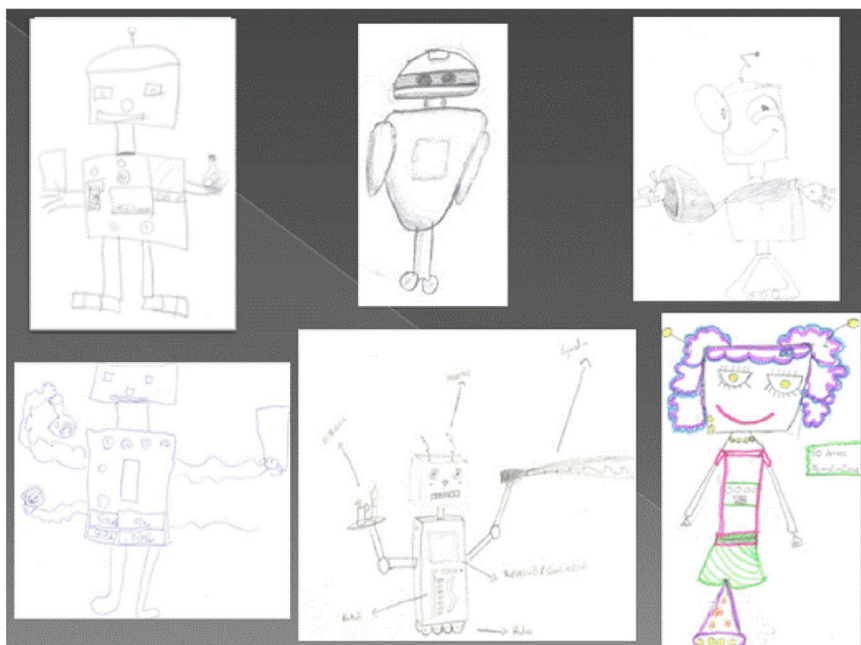


Figure 1: Mental Representations and Expectations in Children's Drawings

Two specific physical features were present in the appearance of most drawings: (i) robots exhibited an interface surface on the chest, tablet-like. According to the inquired the interactive surface on the chest was suitable for ordering a pizza, hamburgers, capable of functioning as an ATM delivering money or just suitable for working as a tablet or console.

(ii) a significant number exhibited several arms suitable for performing different domestic tasks. As the figures above illustrate this multi-functional character associated to the robot was translated in the presence of several arms

Overall robots were expected to act as playmates, help in school tasks and help tidying the room.

<sup>‡‡</sup> For a detailed account of this survey, c.f., Ferreira [2]

### 3.3. The Physical Appearance – From First Drafts to the Definite Image

In this phase were also discussed the basic features of the robot: Its size, weight and mobility and their adequacy to the children and to the spatial context, an environment in which the robot could not become an obstacle to the movement of the staff. In terms of dimensions it was essential to take into account the fact that the children's height could be either that of a toddler or that of a teenager and the importance of attempting to keep "face to face" contact as much as possible. Overall, the robot was planned to be about 120cm, with smooth curved shapes, moving slowly and steadily.

Design cues were captured by the industrial designer IDmind from the children's drawings in order to correspond to some of the referred expectations.

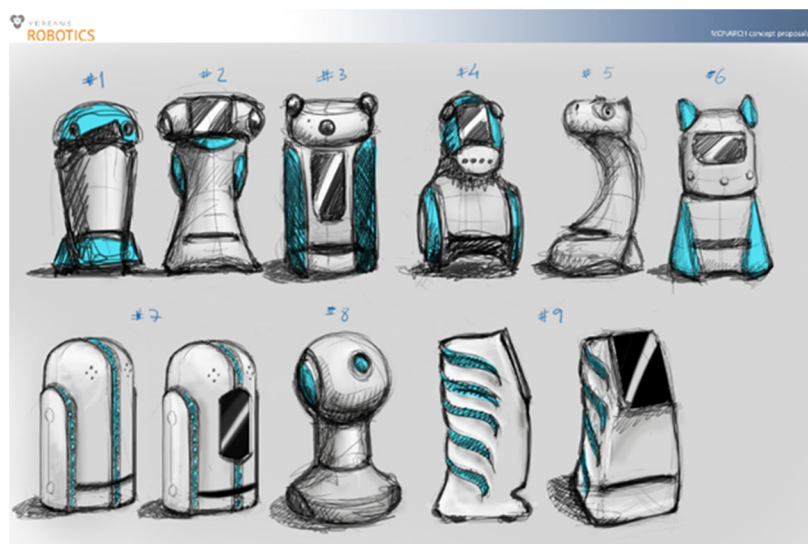


Figure 2: Mbot's first design proposals

It was decided that the following sketches would somehow merge the options #2 and #3 in the figure above.

Safety of the children and bystanders was of paramount importance, and it was addressed directly by the design of the robots, through bumpers and proximity sensors, as well as the soft material substance and curved shapes of the outer shell.

It was also believed that successful interaction heavily depended on the expressive power of the robot. Consequently, the Mbot was equipped with the following expressive and communicative capabilities:

- Friendly facial look resulting from the combination of eyes, cheeks, and mouth expression



- Different mouth shapes displaying (through a LED's matrix) different emotional states: happiness, surprise, sadness, or unpleasantness.
- Variation of the colour of the cheeks and/or the intensity of the light of the eyes to display a specific mood.
- Non-verbal sounds to express attention, angry, amazing, confirmation, error, greetings, warning, or thinking.
- Verbal communication was defined according to Portuguese Text-to-Speech tool.
- A touch screen placed in the mbot at chest level for further input during HRI.

This resulted in a look that was perceived by children as **cute, funny and kind**.



Figure 3: Phases 2 ( on the right) and 3 (on the left)

### 3.4. User's Experience

Before the actual deployment of the robots, some mock ups were placed in the Pediatric Ward to test children's reaction to the robot's image. Children seemed attracted by the picture and found it funny. This was verified by watching their reactions, though this information was not subject of statistical quantification.

The effective deployment of robots looked successful, since children immediately named the robot- Little Casper establishing an affective interaction with it.

The children, however, never assumed the robot as a living entity as their pets are, they were always aware they had to be recharged and that there were people responsible for their maintenance,

The nature of this interaction was assessed throughout 18 months of continuous trials that constituted a long-run experiment described in Ferreira and Sequeira [8]

#### **4. Prospective Insights**

Design has progressively shifted away from being exclusively functionally oriented or even aesthetically driven towards a framework where the complexity of the physical, social, cultural and psychological reality of each individual and their well-being is considered.

A user-centred perspective is particularly relevant in the design of robotic systems intended to share a lived space with people, interacting with them daily.

In fact, successful Human Robot Interaction depends not just on the user's acceptance of the presence of a robot performing efficiently in one's physical and social environment but on the adequacy to the user's reality, e.g., the size of their home, the expected functionalities and roles to be performed....

Though years ago, I advocated the value of the robot's expressiveness and even its affectionate appearance, today I think this expressiveness should not contemplate showing affection as exhibiting "hearts" or simulating "hugs", i.e. behaving according to human emotional standards. I believe these technological artifacts should have a pleasant appearance (after all who wants to have an ugly machine at home, or have to interact with it on a daily basis at the office?) but should not mimic human emotional behaviors.

#### **References:**

- 1.M.I.A. Ferreira, J.S.Sequeira The concept of [robot] in children and teens: some guidelines to the design of social robots. *International Journal of Signs and Semiotic Systems*, 3(2), July-December 2014, special issue on "The Semiosis of Cognition".
- 2..M.I.A. Ferreira. Designing Social Robots: The Role of the User's Referential Framework. *Procs. of the 20<sup>th</sup> International Conference on Climbing and Walking Robots and Support Technologies for Mobile Machines (CLAWAR 2017)*, London. UK.
3. D. Norman and S. Draper, *User-Centered System. New Perspectives on Human-Computer Interaction*" Lawrence Erlbaum Associates, Inc., Publishers (1986)
- 4.T. Winograd, J. Benett, and L. Young. *Bringing Design to Software*. ACM Press 1996
5. ISO9241-210, available at [http://www.iso.org/iso/catalogue\\_detail.htm?csnumber=52075](http://www.iso.org/iso/catalogue_detail.htm?csnumber=52075) (2010).
6. Marc Steen, *Human-Centered Design as a Fragile Encounter*" *Design Issues* 28(1):72-80 (2012).
- 7.. G. Giacomini, *What is human centred design?"* *Proceedings from the 10<sup>o</sup> Congresso Brasileiro de Pesquisa e Desenvolvimento em Design*. So Lus, Maranhão, Brasil (2012).
8. M.I.A. Ferreira I.A. and J. S. Sequeira (2015) "Assessing Human-Robot Interaction: The Role of Long-Run Experiments". *Procs. of the 18<sup>th</sup> International Conference on Climbing and Walking Robots and Support Technologies for Mobile Machines (CLAWAR 2015)*, HangZhou, China, 6-9 September 2015.

## UNAWARE SELF-NUDGING BY SOCIAL ROBOTS

STEFANO CALBOLI

*Centre for Ethics, Politics and Society, University of Minho, Campus de Gualtar,  
Braga, 4710-057, Portugal  
E-mail: calbolistefano@gmail.com*

This work focuses on the use of social robots to perform an understudied form of nudging, namely unaware self-nudging. Equipped with a conceptual analysis of both aware and unaware self-nudging and capitalizing on the literature on the ethics of traditional nudges, the article investigates the ethics of unaware self-nudging through social robots. The gist of the paper is that the chance to personalize the safeguards needed to ethically perform unaware self-nudging through social robots calls for the need for ad hoc regulations on disclaimers to be introduced by social robot suppliers.

### 1. Introduction

Nudges are “any aspect of the choice architecture that predictably alters people’s behaviour without forbidding any options or significantly changing their economic incentives. To count as a mere nudge, the intervention must be easy and cheap to avoid. Nudges are not mandates” [1]

Nowadays, nudges are policy tools that need no introduction. Since the publication of *Nudge: Improving Decisions About Health, Wealth and Happiness* by Thaler and Sunstein [1,2], scholars have widely investigated, on the one side, their effectiveness by means of laboratory and field experiments and, on the other side, their ethical status. Likewise, policymakers wasted no time in introducing policies based on nudges as witnessed by the establishment of nudge units as advisory bodies worldwide.

Let us briefly consider two cases considered paradigmatic examples of nudges. First, the cafeteria case described in Thaler and Sunstein [1]. In the cafeteria example, the canteen manager decides to place the salad at eye level on the counter rather than french fries expecting, in light of cognitive sciences, to increase the probability that the salad is chosen. Second, let us consider a nudge based on, arguably, the effect most successfully exploited by nudgers: the default effect. The default effect is a phenomenon whereby one option in a defined set of options is selected more frequently if the decision-maker ends up with that option doing nothing.

This effect has been successfully exploited in a plethora of cases [3], more famously in the case of organ donation: in the countries in which organ donation is set by default (option to opt-out) donations are remarkably higher than in the countries where the will of donate organs implies an explicit expression (option to opt-in) [4]. These examples, the *cafeteria case* and the *nudging by default* would help the readers in considering the case on focus in the article, namely a special kind of self-nudging, a somehow surprisingly understudied form of nudging.

Self-nudges are those nudges in which the nudger, that is who wants to encourage a certain behavior through an intervention, and the nudgee, namely who gets the behavior influenced by that intervention, are the very same person.

The article aims to show that a special form of self-nudging, albeit less ethically troublesome than traditional nudges is not free from ethical issues.

The following section (§2) is devoted to delimiting the boundaries of a specific kind of nudging, self-nudging and distinguishing between aware and unaware self-nudging. §3 considers the ethics of traditional nudging intending to identify ethical issues applicable even to

cases of unaware self-nudging. The article moves on to recognize the peculiarity of the ethics of unaware self-nudging and discusses the need to comply with the right of awareness of the nudging processes introduced by the technology acquired (§4). The gist of the paper is unfolded in §5 and in the final section. §5 dwells into the practical chance to personalize the releasing of information that must be made available to comply with the consumer's right to awareness. §6 concludes by considering some specific cases of personalization particularly interesting from an ethical viewpoint and proposing ad hoc regulations to make personalizations ethically fully legitimate.

## 2. Aware and unaware self-nudging

A fine-grained analysis of self-nudging is needed to enter the debate on the ethics of self-nudges. First, we should distinguish between two kinds of self-nudging in light of the role played by the nudgee. On one side, we have cases in which the nudgee, in a bid to nudge herself, is the one who identifies and employs the (combination of) nudge(s) meant to promote the targeted behaviors. When so, the nudgee is necessarily aware of the techniques employed. Let us refer to these cases as *aware self-nudging*, namely what is usually considered self-nudging. For example, *aware self-nudging* is the key policy tool in the riveting proposal made by Viale in *Nudging* [5]. Viale considers self-nudging the condition to make nudges factually legitimate in light of libertarian values. Viale [5], primed by Reijula and Hertwig [6], builds an argument for aware self-nudging for which policymakers should make available to decision makers “fact boxes” meant to make them able to adopt nudging strategies to achieve their behavioral aims: “A fact box describing a self-nudge should include five main pieces of information: (1) a description of the behavioral problem; (2) a description of a suitable self-nudge; (3) an explanation of the psychological mechanisms underlying the self-nudge and how it can help to mitigate the problem; (4) an actionable description of how to implement the nudge (if necessary, with links to additional tools and resources; e.g., a hyperlink to an app); and (5) if available, a list of the possible benefits and potential side effects (in terms of easily understandable effect sizes)” [6, p. 25]. In this way, paternalism is straightforwardly avoided being the tools introduced by the nudgees themselves.

Nevertheless, self-nudging can be in place even through different modalities. Indeed, It could be the case that the nudgee, albeit oblivious to nudging techniques, still practices self-nudging: *unaware self-nudging* is too a possibility.

To fully grasp the bit counterintuitive case of unaware self-nudging, let us gradually approach the case, considering first instances of different nature sharing with unaware self-nudging only some aspects. *Self-exclusion from gambling* is a regulation introduced by US policymakers for which gamblers can demand to be denied access to places where they would engage in gambling [7]. In some way, such regulation recalls the story told in the *Odyssey* (XII book) where Ulysses asked his fellow sailors, endowed with beeswax to be plugged up in their ears, to be tied up to the mainmast and ignore his asking to be untied, in order to enjoy safely the sirens' clear-toned song. In both these cases, the decision-maker is aware of her limits and takes precautions beforehand. However, self-constraint is not the only strategy available, decision-makers could indeed take advantage of a softer approach and *nudge* themselves toward the yearned behaviors. This is the idea advocated by Viale [5] and Reijula & Hertwig [6] where the decision-makers *learn* how to nudge themselves. In a way, this is a toned-down version of the *self-exclusion from gambling* and the *Sirens and Ulysses* case just saw. In aware self-nudging, decision-makers leave themselves a way out from the behavior deemed as preferable, indeed, any coercive measures, such as denied access and painters, are involved.

Now let us consider two close scenarios but only one of the two depicts a case where the decision maker nudges herself without knowing anything about nudge techniques. *Caio*, an overweight person, wants to lose weight but struggles to behave accordingly with this desire. He reads the manifesto book of the nudge theory, *Nudge* [1], and knows, from the cafeteria case, the effect of placing the targeted product at eye level. Therefore, he decides to apply the same

technique at home and rearrange the pantry accordingly. In a second scenario, we have *Sempronio*, an overweight man struggling to lose weight. Sempronio did not read the book *Nudge* looking for solutions to improve his life and rightly believes that he could lose weight by training. He has faith in technology and buys a social robot meant to persuade users to train, especially by taking advantage of nudges. In this last scenario, Sempronio is the nudger, in that he is who wants to encourage a certain behavior through an intervention, *and* he is as well the nudgee, namely who gets his behavior influenced by that intervention. This makes the scenario starring Sempronio a case of self-nudging and yet Sempronio is oblivious to the techniques adopted by the social robot to encourage the training: unaware self-nudging is in place.

It must be noted that the scenario starring Sempronio is anything but unlikely. For instance, we can consider the robot trainer developed by Rea, Schneider and Kanda [8]. This robot has been developed to test the efficacy of polite and impolite sentences in motivating users to train at their best. It is easy to imagine a further development given that the robot says nudging sentences. For instance, the robot could take advantage of peer pressure and social norms, phenomena vastly exploited by nudgers [see for instance 9], and device sentences such as “most of the trainers at your age work more and better than that!”.

A further example of a social robot that can be considered an unaware self-nudging technology is a further development of a robot devised by Ali Mehenni and colleagues [10]. These scholars developed a nudger dialogue system, Pepper, aimed to promote altruism among children from five to ten years old. We can easily imagine a close case in which an adult, of his own volition, buys a system like that, calibrated on adult users, after realizing to tend to behave too selfishly.

We considered two examples where the social robot is the technology adopted to perform unaware self-nudging, however, it is not necessarily the case, persuasive technologies can be of different kinds, such as software applications and mobile devices [11]. Nevertheless, social robots represent a technology particularly suitable for unaware self-nudging. The reason is twofold. First, other than nudging users directly, as in the cases of the robot trainer and the nudger dialogue system just considered, social robots can nudge as well indirectly, being able to move around physical choice environments and intervene in them. For instance, a robot could be programmed to rearrange the pantry in light of insights from behavioral and cognitive sciences while mobile devices cannot do that. Second, as noticed by Rodogno [12], social robots are peculiar in the sense that they are particularly suitable to nudge, other than through the standard tools, influencing the overall cognitive and affective users' states.

Social robots are technologies looking particularly suitable to nudging, in general, and to the practice of unaware self-nudging, in particular. This makes it urgent to debate the ethics of unaware self-nudging through social robots, as we will see a case rich with aspects ethically relevant. Now that we have outlined what unaware self-nudging is, we are ready to investigate ethical issues emerging from unaware self-nudging through social robots.

### 3. Ethics of self-nudging

Concerning unaware self-nudging through social robots, are there forgotten ethical issues? To be deemed as a legitimate practice of persuasion, are there precautions that should be in place? Does the use of social robots open peculiar circumstances ethically relevant?

Before spreading social robots performing unaware self-nudging, it is advisable to attempt to answer these questions. In this regard, the analysis ethicists are asked to perform cannot be limited to analyzing the ethically relevant consequences of the introduction of a technology. Rather, ethicists should be involved from the very beginning in the development of technologies and promptly identify precautions to improve their ethical status. In doing so, ethicists act in advance and rather than considering exclusively scenarios that *necessarily* emerge, should as well examine *possible* scenarios that could or could not emerge in light of (among other things)

the specific users' needs, society's reactions and suppliers' advantages. Possible scenarios should be part of the ethicists' agenda as much as necessary scenarios.

The investigation of possible scenarios emerging from unaware self-nudging through social robots is the focus of the paper. Specifically, we will consider how a specific feature of social robots that represents a marketing advantage opens scenarios worth to be explored.

With this purpose, we begin by taking advantage of the extensive literature on the ethics of nudges, consider the ethical issues raised when traditional nudges are on focus, and see if they are relevant for unaware self-nudging through social robots.

The ethical issues typically raised when nudges are on focus are essentially two. The first one concerns decisional autonomy, which often is meant as the ability to deliberate on decisions and behaviors [see, for instance, 13]. Arguably, there are at least some cases in which nudges could impair decisional autonomy, namely the cases where decision-makers would have relied on deliberation and chose *A* but, due to the nudge, in fact, chose *B*. For instance, let us consider a scenario where the nudger employs a nudge based on the default effect, as in the case of organ donations seen in the first section. A hypothetical decision maker, Tizio, could be hesitant about what to do, to donate organs or otherwise. Let us suppose that, if asked directly, Tizio would have weighed the pros and cons, deliberated and preferred to not donate. However, the policymaker opted for setting organ donation by default and our nudge, Tizio, allured by the effortless, unconscious and intuitive cognitive processes triggered by the nudge, ended up donating organs. It seems that, in this case, the decisional autonomy of Tizio has been violated. Nevertheless, a convincing reply can be made. Even this kind of case looks unproblematic provided that there are some strategies available to nudgees to resist the nudge and, usually, there are. For instance, consider the case in which a would-be nudgee, enters a supermarket where some products are displayed on eye level to increase the probability of being purchased. If wanted, customers can beforehand arm themselves with a shopping list and buy only what is on the list, regardless of where products are placed on the supermarket shelves. Typically, would-be nudgees are in the position to decide in advance how to behave in the future, regardless of choice environments' traits. For the purpose of this work, it is immaterial trying to identify cases in which it is particularly hard for would-be nudgees to take precautionary measures. Instead, our focus is on unaware self-nudging through social robots, and this specific case is not particularly problematic in that self-nudgees can always dispose of the technology should be deemed necessary.

Nevertheless, it should be noted that especially considering multi-purpose technologies as social robots, ethical and pragmatic aspects are mixed, and it would look desirable to give purchasers the chance to reprogram the social robot and employ it exclusively for other purposes rather than dispose of them. Hopefully, this function will represent an alluring market advantage for social robot suppliers. Relevant here, regardless if such function is guaranteed or otherwise, decisional autonomy is anyway safeguarded.

Albeit probably the most discussed, decisional autonomy is only one of the two ethical issues examined when traditional nudging is on focus. The other ethical issue concerns public scrutiny [see 13, 14]. Nudgee should be made able to scrutinize and evaluate the work of policy-makers/nudgees, this is a crucial aspect in the unfolding of democratic processes in our modern liberal democracies. To see why guaranteeing decisional autonomy is insufficient to make the public scrutiny of nudges possible, let us consider, once again, the case of organ donations. A decision-maker could discuss with family and peers, reading up and reaching the conclusion that not donating organs is the right choice. If so, the decision maker would inquire about the modality to express her will and act accordingly. Hence, the nudge does not impair the decisional autonomy of the decision-maker. However, we should ask ourselves, is the decision maker able to spot the nudge? The answer is no in that the fact that organ donation is the default option could have been as well an accidental trait of the choice environment. The decision maker cannot distinguish between the two possibilities and, as a result, be able to evaluate the policy-making, its strengths and the fact that the policymakers preferred that nudge over alternative policy tools equally available. In view of all this, making transparent the presence of the nudge looks necessary, transparency indeed makes distinguishing a nudge from an accidental trait of

the choice environment possible, and, in turn, scrutinizes the work realized by policymakers. However, public scrutiny is not an issue when nudges leave the public sphere and enter the private sphere. When a person purchases a social robot meant to perform unaware self-nudging, it is completely irrelevant to the unfolding of democratic processes that that purchaser is able to detect the nudges adopted being them tools to steer, in a targeted manner, her behavior rather than a policy impacting on all the citizens.

Summing up, concerning decisional autonomy, unaware self-nudging by social robots seems to not pose any intricacy, and the issue of public scrutiny does not even apply to the case. So, should we conclude that the particular case of self-nudging at hand is unproblematic? Should we conclude that ethicists could do nothing but recognize the legitimacy of such kinds of nudging through social robots? In the next section, I argue that, as a matter of fact, the case at hand cannot be set aside by ethicists, especially because of a scenario made possible by a peculiar trait shared by many technologies, social robots included.

#### **4. The ethics of unaware self-nudging by social robots**

As already mentioned, in modern liberal democracies decisional autonomy is highly valued. The pivotal role of decisional autonomy in our democracies does not emerge exclusively when the public sphere is considered, rather, it plays a role even in the private sphere, when the relationships between consumers and suppliers are in place. When a product or a service is offered on the market, the right of the consumers to be aware of their features should be ensured. Let us call it the *right of awareness of the product*. European law cannot be any clearer on the necessity to ensure this right. In the fifth article of the *Directive 2011/83/EU of the European Parliament and of the Council* is stated that “the trader shall provide the consumer with the following information in a clear and comprehensible manner if that information is not already apparent from the context [...]: *the main characteristics of the goods or services*, to the extent appropriate to the medium and to the goods or services” (italics mine).

The function of social robots conceived to self-nudge is to shape the choice environment in a way for which the chance to behave as desired is increased. It follows that among the characteristics on which the user of a social robot performing unaware self-nudging must be informed, there are nudges, namely the means through which the technology performs its function. It is so in that it is hard to argue that the nudges are “information already apparent from the context”, namely, as specified by the *Guidance on the interpretation and application of Directive 2011/83/EU of the European Parliament and of the Council on Consumer Rights*, self-evident information, such as the trader’s geographical address and identity when contracts other than distance or off-premises are in place. Hence, to comply with European legislation, the user should be aware that the nudge adopted by social robots.

What, contrariwise, does not look straightforward is the level of depth on which information on nudges should be released. However, even in this respect, we can keep referring to the European legislation on consumers’ rights and try to build an analogy with a case sharing many similarities with the case at hand: online research queries. In particular, we can refer to the *Proposal for a Directive of the European Parliament and of the Council amending Council Directive 93/13/EEC of 5 April 1993, Directive 98/6/EC of the European Parliament and of the Council, Directive 2005/29/EC of the European Parliament and of the Council and Directive 2011/83/EU of the European Parliament and of the Council* as regards better enforcement and modernisation of EU consumer protection rules. In this proposal amending, the cases in which the trader provides information through search results in response to the consumer’s online search are considered. As regards the information required regarding the parameters determining the ranking displayed, the XIX point of the proposal claims that “traders should not be required to disclose the *detailed functioning* of their ranking mechanisms, including algorithms. Traders should provide a *general description of the main parameters* determining the ranking that explains the default main parameters used by the trader and their relative

importance as opposed to other parameters” (Italics mine). By way of analogy, when social robots performing unaware self-nudging are on focus, this level of informational depth would correspond to inform the user of the kind of nudges the robot is able to introduce, rather than go into the detail on the effects exploited by those nudges or even on the mechanisms underpinning these effects [on the difference between effects and mechanisms see 15,16]. Such kind of information should be made available to the purchasers of social robots performing unaware self-nudging to comply with the right of awareness of the product. In a sense, transparency on nudging processes, namely the safeguard to decisional autonomy in traditional nudging, is still due but for different reasons. Unaware self-nudging needs the introduction of some precautions to be considered ethically legitimate. In the next sections, we further see how the request for information opens a possible scenario made possible by the nature of the technology under consideration.

## 5. Personalization in information releasing

As pointed out in the second section, ethicists should as well deal with possible scenarios emerging from the use of technologies. In this respect, there are aspects ethically relevant to ensuring the right of awareness when social robots performing unaware self-nudging are purchased.

Social robots are particularly alluring for both suppliers and users due to their feature to be easily personalizable according to the user’s need. It is vastly recognized in the literature that personalization helps to optimize the acceptability and willingness of a specific user to engage with social robots [see 17].

Social robots are technologies meant to interact with users for an extended period of time and personalization is key to maintaining user engagement and promoting trust toward the robot [see 18]. Furthermore, particularly relevant for our considerations, personalization looks like an unmissable opportunity for social roboticists with respect to individual preferences on privacy [19, see also 20]. Personalization looks, other than a feasible and advantageous practice, a captivating opportunity from the suppliers’ perspective, who can ameliorate the technology put on the market and the user’s perspective, who can set the robot’s behaviours according to her specific needs.

In light of the alluring of personalizing social robots, it looks like a possible scenario, and arguably a predictable one, that users will have the opportunity to personalize the ethical safeguards to nudging (on personalization applied to nudges’ transparency see 21). Relevant here, social robots assure the chance to personalize the release and the degree of accuracy of information. Evidently, personalization cannot disregard legislative provisions and so it should guarantee at least the kind of information requested by them. However, it still can be personalized the release of information further increasing the degree of detail and as well intervening in the aspects on which the law is salient.

To get started on what the personalization of information could look like, information can be personalized in terms of time release. Users could be enabled to choose to get information only at the moment of purchasing, or rather during the use of the social robot with a certain periodicity. Further types of personalization could concern the contents of the information itself. First, social robots can be programmed to either release general information on the nudges the robot is *able to adopt* or, more accurately, information on the *exact nudge employed* during its activity within the choice environment inhabited by the user. For instance, the user could be made aware of the fact that the robot can move objects to make it more or less probable that they are picked or, in greater detail, the user can get the list with the exact intervention introduced by the robot, for instance, the moving of sugary drinks in a spot of the pantry particularly hard to access and of vegetable at eye-level. Second, personalization could concern the degree of detail of the information released on nudges. We could consider the case for which social robot suppliers should, by law, inform users of the kind of interventions on the choice environment the robot is able to introduce. However, the degree of detail could be increased and, as a result, users can be informed on the effect exploited by the nudge. Hence, the



information requested to comply with the right of awareness of the product could be personalized when social robots are employed to perform unaware self-nudging. This opportunity opens an ethically relevant scenario. In the next and conclusive section, this scenario is discussed.

## 6. Conclusion

Personalizations of information reveal uncharted spaces where empirical questions are intertwined with normative considerations. First, it is urgent to begin the exploration of how the interaction between the release of relevant information and nudges affects the strength of the latter. Concerning the effect of transparency, namely a safeguard relevant to decisional autonomy, on nudge's strength empirical evidence is mixed but certainly, we cannot claim that given *any case* transparency does not impair the strength of nudges [see 22,23]. This suggests that we should investigate if there are undesirable cases concerning information release where some forms of personalized information drastically reduce nudges' strength. Unfortunately, we cannot completely rely on the evidence so far collected concerning the transparency of nudge in that it regards a safeguard introduced for reasons different than those motivating the releasing of information on the product.

Should we identify combinations between information and nudge particularly worrisome, it will look advisable that policymakers set ad hoc regulation asking social robot suppliers to provide disclaimers meant to apprise users about the *consequences* of the personalization preferred in terms of nudges' strength. For instance, let us consider a case in which the user opts for a high degree of detail and as well for the release of such information at the exact moment the nudges are introduced. It could be well the case that this kind of personalization makes interactions with the social robot boring and distasteful, breaking the harmony of the interaction and, as a result, hampering the nudging processes.

It seems reasonable to provide that, once data is collected and confirmed that a specific kind of personalization is particularly worrisome, the user who prefers such kind of personalization is alerted. This warning should consist in a disclaimer where it is reported that data shows how the kind of detail and release of information selected could get the user away from the behavioural aim longed for. This kind of disclaimer would be meant to make users perfectly aware of the consequences of personalization when interacting with social robots, in line with the right to awareness of the product. Requesting suppliers to introduce these kinds of disclaimers should not look exceeding or absurd being a common practice in different contexts. For instance, thanks to the *Patient Protection and Affordable Care Act (ACA)*, at least since 2018 big US restaurant chains are mandated to release calorie information (aka health consequences of eating certain food) concerning menu items.\*

Finally, let us see how the approach just outlined, hence an approach meant to further raise awareness in employing social robots applies as well to an extreme case of personalization. Let us consider once again the case of Caio, an overweight man who wants to lose weight and buys a social robot performing unaware self-nudging to succeed. Let us suppose that Caio believes, rightly or wrongly, that to accomplish his aim it would be ideal to renounce once and for all, irrevocably, the right to be aware of social robots' features, especially of the nudges that the technology can adopt. We can imagine, for instance, that Caio believes that detecting the nudges would lead him to admit his *akrasia* and, in turn, surrender to his weakness. Should this kind of extreme personalization be permitted? I think the answer is affirmative and the reason is twofold. First, this irrevocable personalization consists in a renouncement of a right limited to a precise context, that is the use of *that* social robot, rather than a general waiver concerning the

---

\* Although it should be noted that “in part, due to several delays in implementation of the law, many large restaurant chains began voluntarily posting calories on their menus before it was required, including McDonald’s, which began labeling in September 2012” [24, p. 1].

chance to be informed on the relevant aspects of *all* tools meant to nudge. Second, the renouncement of the right is, in some relevant sense, factually always revocable. Indeed, the user, being the owner, is always able to dispose at will of the technology. Nevertheless, if the aim of the regulation, as suggested by the EU regulation, is to make users more aware of the good or service purchased, it would be advisable to set a regulation to notify users, provided that this is supported by data, that the choice to renounce the right of awareness of the product amounts to a suboptimal personalization in terms of nudges' strengths.

To conclude, the considerations drawn in this article are in some sense the natural continuation of the proposal advanced by Viale to use “fact boxes” to make users able to adopt nudging strategies to achieve the behavioural aim they identified [5]. Even if, here, we are dealing with cases in which users are not familiar with nudge theory.

I hope to have shown how considering practical possibilities opened by a specific kind of technology, namely social robots, in parallel with normative considerations unveils possible scenarios concerning personalizations of information that ethicists cannot overlook.

### Acknowledgements

This work has been supported by *Fundação para a Ciência e a Tecnologia*, prot. UI/BD/152568/2022.

### References

1. R.H. Thaler and C.R. Sunstein, *Nudge: Improving Decisions About Health, Wealth, and Happiness*. New Haven: Yale University Press (2008)
2. R.H. Thaler and C.R. Sunstein, *Nudge: The Final Edition*. London: Allen Lane Penguin Books (2021)
3. J.M. Jachimowicz, S. Duncan, E.U. Weber and E.J. Johnson, When and why defaults influence decisions: A meta-analysis of default effects, *Behavioural Public Policy* **3**(02), pp. 159–16 (2019)
4. E.J. Johnson and D. Goldstein, Do defaults save lives?, *Science*, **302**(5649), pp. 1338–1339 (2003)
5. R. Viale, *Nudging*. Cambridge, MA: The MIT Press (2022).
6. S. Reijula and R. Hertwig, Self-nudging and the citizen choice architect, *Behavioural Public Policy* **6**(1), pp. 119 - 149 (2019)
7. L. Kraus *et al.*, Self-exclusion from gambling: A toothless tiger?, *Front. in Psychiatry*, **13** (2022)
8. D.J. Rea, S. Schneider, and T. Kanda, Is this all you can do? Harder!: the effects of (Im)polite robot encouragement on exercise effort, *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction* [Preprint], pp 225–233 (2021)
9. C. Bicchieri, *Norms in the wild: How to diagnose, measure, and change social norms*. New York, NY: Oxford University Press (2017)
10. H. Ali Mehenni, S. Kobylanskaya, I. Vasilescu, and L. Devillers, Nudges with conversational agents and Social Robots: A first experiment with children at a primary school, *Lecture Notes in Electrical Engineering*, pp. 257–270. (2020)
11. B.J. Fogg, *Persuasive technology: Using computers to change what we think and do*. Amsterdam: Morgan Kaufmann (2011)
12. R. Rodogno, Nudging by Social Robots, *Frontiers in Artificial Intelligence and Applications*, **335**, pp. 337-345(2020)
13. A. Schmidt, The Power to Nudge, *American Political Science Review* **111**(2), pp. 404–417 (2017).
14. A. Alemanno and A. Spina, Nudging legally: On the checks and balances of behavioral regulation, *International Journal of Constitutional Law*, **12**(2), pp. 429–456 (2014)

15. S. Calboli and V. Fano, Mechanistic explanations and the ethics of nudging, *Rivista Internazionale di Filosofia e Psicologia*, **13**(3), pp.175-186. (2022)
16. T. Grüne-Yanoff, Why behavioural policy needs mechanistic evidence, *Economics and Philosophy*, **32**(3), pp. 463–483. (2015)
17. N. Gasteiger, M. Hellou, and H.S. Ahn, Factors for personalization and localization to optimize human–robot interaction: A literature review, *International Journal of Social Robotics*, **15**(4), pp. 689–701 (2021)
18. L. Tian and S. Oviatt, A taxonomy of social errors in human-robot interaction, *ACM Transactions on Human-Robot Interaction*, **10**(2), pp. 1–32 (2021)
19. M. Dietrich, Towards Privacy-preserving Personalized Social Robots By Enabling Dynamic Boundary Management, *Proceedings of the Workshop on Personalization in Long-Term Human-Robot Interaction at the 2019 International Conference on Human-Robot Interaction* (2019)
20. H. W. Park, I. Grover, S. Spaulding, L. Gomez, and C. Breazeal, A Model-Free Affective Reinforcement Learning Approach to Personalization of an Autonomous Social Robot Companion for Early Literacy Education, *AAAI*, **33**(1), pp. 687-694 (2019)
21. S. Calboli, Robot Nudgers. What About Transparency? *Software Engineering and Formal Methods. SEFM 2022 Collocated Workshops, Lecture Notes in Computer Science*, vol. 13765, Springer International Publishing In P. Masci, C. Bernardeschi, P. Graziani, M. Koddenbrock, M. Palmieri (Eds.) (2023)
22. E. Kantorowicz-Reznichenko, and J. Kantorowicz, To follow or not to follow the herd? transparency and social norm nudges, *Kyklos*, **74**(3), pp. 362–377, 2021
23. H. Bruns, E. Kantorowicz-Reznichenko, K. Klement, M. Luistro Jonsson, and B. Rahali, Can nudges be transparent and yet effective?, *Journal of Economic Psychology*, **65**, pp. 41-59 (2016)
24. J. Petimar, M. Ramirez, S.L. Rifas-Shiman, S. Linakis, J. Mullen, C.A. Roberto and P. Block, Evaluation of the impact of calorie labeling on McDonald’s Restaurant Menus: A natural experiment, *International Journal of Behavioral Nutrition and Physical Activity*, **16**(1) (2019)

## **SOCK IT TO CHALLENGING BEHAVIOUR - DEVELOPMENT OF SMARTSOCKS TECHNOLOGY FOR EARLY DETECTION OF DISTRESS IN PEOPLE WITH NEUROLOGICAL DISORDERS: A SURVEY STUDY**

IVETA EIMONTAITE<sup>1</sup>, ZEKE STEER<sup>2</sup>, JACQUI ARNOLD<sup>2</sup> and PATRICK OGUNDELE<sup>2</sup>

<sup>1</sup>*School of Aerospace, Transport and Manufacturing, Cranfield University, College Road, Cranfield,  
MK43 0AL, United Kingdom*

*E-mail: [Iveta.Eimontaite@cranfield.ac.uk](mailto:Iveta.Eimontaite@cranfield.ac.uk).*

*[www.cranfield.ac.uk](http://www.cranfield.ac.uk).*

<sup>2</sup>*Milbotix Ltd, Oxfordshire, OX7 5JG, United Kingdom*

*E-mail: [Zeke@milbotix.com](mailto:Zeke@milbotix.com), [Jacqui@milbotix.com](mailto:Jacqui@milbotix.com), [Patrick@milbotix.com](mailto:Patrick@milbotix.com)*

*[www.milbotix.com](http://www.milbotix.com)*

Increasing life expectancy and healthcare standards are among several reasons for the increase in the aged population across the Western world. However, the impact of this trend means more people will rely on healthcare services, placing a strain on an already limited resource. Experienced feelings of distress and agitation are named as significant factors affecting individuals' wellbeing, access to healthcare, and the impact of care services among people living with dementia and autism spectrum disorder (ASD) populations. Early indication of distress via SmartSock wearables could alleviate the strain on healthcare systems, as well as improve the wellbeing of the technology users and their carers. The current study explored the impact, causes, and solutions to feelings of distress among 101 participants including carers, people living with dementia, autism, learning disabilities, and cognitive impairments via an online survey. The importance the technology was compared among the participant groups. The results indicate that carers and the cognitive impairments group showed significant differences in the usefulness of "help to identify sources of distress though alert system" and "an early indication of pain of anxiety" functions. The study provides evidence on the impact of distress and the expressed need for solutions in detecting it earlier. It also starts the dialogue and collaboration with the target user groups to conduct a feasibility study of SmartSocks technology usage in a care home environment.

### **1. Introduction**

The life expectancy in the Western world has increased over the last several decades due to improvements in living standards and healthcare. However, the aging population poses several challenges for the future, such as increased pressure on healthcare systems (Kojima et al., 2019). Social care in the UK is already burdened by budget cuts (The Kings Fund, 2019), rising labour costs, staffing difficulties (Skills for Care, 2022), and the aftermath of COVID-19. This emphasizes the need for smart solutions to ensure the well-being of people in care and create a less stressful environment for staff. Furthermore, the need for technological solutions is underscored by the predicted drastic increase in age-related health conditions over the next few years. For instance, the number of people living with dementia in the UK is projected to double to 1.6 million over the next two decades (Alzheimer's Research UK, 2023), and 2% of children and adults in the UK are already affected by autism. These examples highlight the growing need to provide solutions that promote and prolong independence for all members of society.

The current study focuses on two main participant populations: people living with dementia and adults with autism spectrum disorder (ASD) and learning disabilities (LD). Over 70% of people with dementia display aggression, irritability, and resistance to care (Ringman & Schneider, 2019). These distress experiences are major sources of burden for carers (Tsai et al., 2021), consuming 28% of their time (Beeri et al., 2002). Additionally, feelings of distress in people living with dementia are predictors of entry into long-term care (Schulz et al., 2004) and are associated with worse health outcomes for both individuals with dementia and their carers, including higher rates of physical and mental illness (Serrano-Aguilar et al., 2006), reduced

quality of life (Molyneux et al., 2008), and increased morbidity and mortality rates (Aneshensel et al., 2000). These distress-related issues are extremely costly, with individual care costs increasing by up to £30,000 per year (Livingston et al., 2014) and adding £2 billion to the annual cost of dementia care in the UK (Morris et al., 2015).

Autism Spectrum Disorders (ASD) is another condition where experienced distress and agitation, known as meltdowns (intense responses to overwhelming situations), can have a detrimental effect on individuals and their carers' well-being. Meltdowns make accessing healthcare difficult, contributing to health inequalities (Weir et al., 2022). Healthcare access for autistic children and young people, including visiting GPs, attending hospital appointments, or dental care, typically involves multiple triggers, and meltdowns serve as significant barriers to accessing care, leading to health inequalities. In addition to being stressful for the autistic child or young person, meltdowns are one of the primary sources of stress for their parents/carers (Broady et al., 2017).

Recognizing agitation early and intervening is challenging but essential to minimize the effects of distress and meltdowns. Once an autistic child or young person is having a meltdown, there is little anyone can do to help except provide them with time and space to recover (<https://www.autism.org.uk/advice-and-guidance/topics/behaviour/meltdowns/all-audiences>). Considering that meltdowns hinder healthcare access (Weir et al., 2022), prolonged agitation over time is likely to result in reduced healthcare access and increased strain on the healthcare system. For people living with dementia, failure to treat agitation early leads to worse clinical outcomes (Kovach et al., 2005). In care homes, survey instruments used to assess agitation are time-consuming (Lai, 2014), often missing early signs of agitation (Kang et al., 2004). In the community, family carers often recognize early signs of agitation but lack the resources to manage it effectively (Herron & Wrathall, 2018). Non-pharmacological interventions like music, reminiscence, and sensory therapies can reduce agitation (Cohen-Mansfield, 2001), but they must be administered before behaviour escalates (Kolanowski et al., 2009).

Agitation has significant effects on the quality of life for individuals themselves, as well as for caregivers and those around them, as discussed earlier. Over the last decade, technological solutions have been investigated to help individuals cope with feelings of distress and agitation. The most common technologies can be categorized as smart home technology (Amiribesheli & Bouchachia, 2018), assistive robotic technology, therapeutic technologies (such as virtual/augmented reality), and caregiver support technologies, including smart photo albums and tablets (Moyle, 2019). While smart home and assistive robotic technology have proven to be effective, they also have their drawbacks (Moyle et al., 2021). Adapting existing infrastructure to integrate such smart technology is not always feasible or cost-effective, and proving efficacy can be challenging. Caregiver support and early intervention technology, such as wearable sensors, are showing the most potential due to their affordable prices and low impact on the existing home environment for integration and use.

The increasing availability of wearable physiological sensors and controlled environment studies showing that physiological data can reliably predict meltdowns (Goodwin et al., 2019) indicate that early intervention is possible. However, the challenge lies in the acceptance of wearables by the target population (Koumpouros & Kafazis, 2019). Sensors have a greater chance of being accepted and used by the ASD population if they blend in with everyday objects and are gradually introduced, allowing for decentralised integration with new objects (Sharmin et al., 2018). Furthermore, an observational study of children with ASD indicated that sensors worn on their bottom calves (Hedman et al., 2012) had the least interference with everyday activities. These findings align with studies examining wearable technology with people living with dementia and their carers (Bankole et al., 2011). Qualitative interviews with nine people living

with dementia and their carers regarding wrist-worn devices (Fitbit) after a physical exercise intervention study (O'Sullivan et al., 2023) revealed challenges in effective use without support from researchers, resulting in participants showing little interest in wearing them after the study. Further exploration using the Unified Theory of Acceptance and Use of Technology highlighted the significant effect of social influence and effort expectancy factors (Dai et al., 2020).

The evidence from usability and user experience studies mentioned above underscores the need for seamless integration of wearable technology and ease of use for the target users. Integrating physiological sensors into everyday garments, such as socks, worn by a large proportion of the population, offers a solution to the acceptance issues discussed earlier. Socks are non-stigmatizing and easily integrate into the daily lives of those wearing them. Furthermore, reliable smart sensors would provide greater opportunities for advancing smart home technologies and promoting more inclusive and supported living for individuals with varying needs. The focus of the current paper is SmartSock technology by Milbotix Ltd integrating heart rate, EDA, skin temperature and accelerometer data to provide early indications of rising distress and agitation. This technology is designed to provide psychological or physical assistance at an earlier stage for feelings of distress and agitation management than otherwise able.

The current study is part of a project that explores the usability and effectiveness of SmartSocks technology. This data represents the second step in a four-step approach involving co-creation engagements with users and technology development stages: (i) focus groups and interviews, (ii) online survey, (iii) feasibility trial with care homes, and (iv) result dissemination and further feedback collection through focus groups. The user engagement and data collection presented in this paper were gathered through an online survey, which focused on exploring the impact, causes, and solutions related to feelings of distress, as well as the desired functionality of smart wearable technology. The survey primarily targeted individuals living with dementia, autism, and learning disabilities. In order to provide a comprehensive perspective on these two populations, a group of carers was also included. Additionally, individuals with cognitive impairments were included as an additional group of interest, aiming to explore the potential for wider application.

## **2. Method**

### **2.1. *Participants***

A total of 113 participants completed the survey on the Qualtrics platform. However, 12 participants were excluded from the final analysis because they did not complete one of the three survey blocks. Therefore, the final analysis included 101 participants. The recruitment of participants was carried out through various channels, including Cranfield University intranet announcements, contacts within collaborating care homes and charities, and the Prolific platform. The majority of participants (90%) were recruited through the Prolific platform. The selection criteria for participants were as follows: (i) carers of individuals living with dementia and/or learning disabilities (including family carers), (ii) individuals living with mild dementia, (iii) individuals with learning disabilities and autism, and (iv) individuals with cognitive impairments (such as mood disorders and anxiety). The demographic information of the participants is provided in Table 1. The study received approval from the Cranfield University Ethics Committee.

## **3. Procedure**

The study was piloted with three adult participants without cognitive impairments and internally reviewed by two subject experts experienced in formal and informal care for individuals living

with dementia and learning/physical disabilities. Some of the question wording, particularly in sections relating to experienced distress, was adjusted to align with conventions and avoid emotional loading.

The survey was conducted online via the Qualtrics platform from February 1st to March 1st, 2023. On average, participants took 10 minutes to complete the survey. There were two versions of the survey, depending on whether participants identified themselves as carers or individuals living with a specific condition. Both versions contained the same questions, but they were phrased differently for carers to discuss the person with the condition. Additionally, in the distress impact block, carers were asked about the impact of the person's distress on themselves as caregivers. Regardless of whether participants were carers or individuals with a condition, they received the same initial information: an outline of the study's aims, participant rights, and data usage, storage, and disposal. The survey then proceeded to demographic questions, with the final question inquiring about different symptoms of distress (such as irritability, anxiety, etc., as outlined in the materials section). If participants indicated that they did not experience distress, they were directed to the end of the survey.

The second block of the survey addressed the causes, symptoms, and management of distress, while the final block explored the evaluation and potential use of SmartSocks as a solution for early detection of distress. This block began with a brief introduction to SmartSocks, followed by several pictures of the technology, and concluded with questions seeking participants' opinions on the functionality of this type of wearable. After completing the survey, participants were presented with a debrief. The study did not contain any forced response questions, except for the informed consent section where participants had to provide consent to participate in the study. If participants did not provide consent, they were directed to the debrief section.

Table 1: participants descriptive information as a function of group

		Carers	Autism/ learning disabilities	Other cognitive impairments	Dementia
Gender	Male/Female/Other	8/41/0	8/13/1	11/9/0	5/5/0
Age	18 - 24	2	5	7	0
	25 - 34	16	2	4	0
	35 - 44	15	10	6	0
	45 - 54	7	2	2	2
	55 - 64	7	3	1	3
	65 - 74	0	0	0	4
	75 - 84	2	0	0	1
Ethnicity	White	42	19	17	9
	Mixed/Multiple ethnic groups	4	1	1	0
	Asian/Asian British	1	1	1	1
	Black/African/Caribbean/Black British	1	1	1	0

### 3.1. *Materials*

The survey comprised three main blocks: Block 1 - demographic information; Block 2 - causes, impact, and coping strategies for distress; and Block 3 - an overview and assessment of the usefulness of the SmartSocks. The questions in Blocks 2 and 3 were derived from previous focus groups and interviews, which are not the primary focus of this paper.

Block 1 consisted of questions related to participants' gender, age, ethnicity, whether they or the person they care for wear socks, whether they have any conditions affecting their feet, and whether they experience symptoms of distress (such as irritability, frustration, being short-tempered, anxiety, stress, feeling overwhelmed, or none of the above). Participants were instructed to select all relevant symptoms from the provided list (multiple selections were allowed). The number of distress symptoms was calculated as a count for further analysis.

Block 2 consisted of three quantitative questions: participants were asked to indicate the severity or level of distress/anxiety on a scale from 1 to 10 (1 representing minor anxiety or no anxiety, and 10 representing the highest possible level of anxiety); participants were also asked to indicate how well they were able to identify the causes of their anxiety/distress (ranging from 1 - All or most of the time to 5 - Never or hardly ever); and participants were asked to rate the extent to which their coping strategies were successful (1 - All the time to 5 - Never). Additionally, participants were asked to describe the impact of distress on themselves and those around them, as well as the observed causes of distress and the strategies they employed to cope with and minimize its impact.

Block 3 started with information about the SmartSocks as a wearable solution indicating rising feelings of distress “*The SmartSocks™ contain electronic sensors that record the wearer’s pulse, skin temperature, sweat, and movement. This information can be used to:*

- *Identify and avoid sources of distress in the environment.*
- *Alert carers to signs of distress so that they can intervene at an earlier stage.*
- *Evaluate interventions for behavioural symptoms and assess how well they are working.*

*Research and development of SmartSocks™ is supported and endorsed by Alzheimer’s Society, and collaborators include the UK Dementia Research Institute and six Universities.”*



Figure 1: SmartSocks being fitted to a person living with dementia\*

Following the previous information, participants were asked about the usefulness of this technology to them, with response options including "Yes, would be useful now," "useful in the future," or "Would not be useful." They were then presented with a question regarding various functions of the SmartSocks, such as "early indication of pain or anxiety," "identification of sources of distress," "understanding day-to-day levels of distress," and "evidence of distress to provide to health services." An open-ended option was also provided for participants to provide their own response. However, participants did not provide input that fell outside of the categories already provided. The final two questions inquired whether participants used smartphones and if they had access to Wi-Fi in the location where they spent the majority of their time. These two questions were not discussed in the results section of this paper.

---

\* [Communities where older people flourish - St Monica Trust](#)



## 4. Results

Two demographic information questions were asked to confirm the suitability of SmartSocks as feasible wearable technology. Participants were asked about the person with the condition and their usage of socks. The results showed that over 80% of participants indicated that they or the person they care for wear socks "all the time" or "most of the time." Additionally, 18 participants reported having "frequent plasters or sores" on their feet. These findings suggest that socks as a smart wearable technology would be acceptable and usable for the majority of the target population.

### 4.1. *Feelings of distress experiences*

Block 2 of the survey focused on exploring participants' experiences and coping strategies for feelings of distress. The differences between participant groups were analysed using the non-parametric Kruskal-Wallis test. However, none of the questions in this block yielded significant differences. This indicates that the impact of distress, as well as the ability to predict causes and the effectiveness of coping strategies, were similar among the different participant groups. Therefore, these aspects were further explored descriptively. The combination of quantitative and qualitative questions allowed for an investigation into the impacts, observed causes, and discovered coping strategies related to distress. Survey participants indicated that they were all familiar with feelings of distress, with frustration and anxiety being the most common (Table 2).

Table 2: Experienced components of distress as a function of participant group (%)

Do you (if not carer)/the person you care for (if carer) experience any of the following?				
	Carers	Autism/ learning disabilities	Dementia	Cognitive impairments
irritability	76	33	50	64
frustration	76	33	70	68
short-tempered	65	23	60	27
anxiety	67	53	90	86
stress	59	37	70	77
feeling overwhelmed	57	43	60	77

Following this, participants were asked to explore the causes of distress. Most participants reported that they could identify the causes of distress "mostly, but not all the time," or "some of the time," with the percentage of responses ranging between 40% and 60% of respondents, and 30% and 47%, respectively. Interestingly, the option "never or hardly ever" was not selected by the survey participants, indicating that having a certain level of understanding of the causes of distress would allow participants, at least to some extent, to manage it (Table 3).

The open-ended question asking participants to indicate how distress affects them illustrates the isolating nature of these feelings. Some participants expressed that distress prevents them from leaving their houses alone and can make them feel sick, tired, and completely overwhelmed. They mentioned withdrawing from certain activities and trying to avoid specific places or events. Others noted that the impact of their distress extends to those around them, with some admitting to verbally lashing out at others during episodes of distress. This can lead to deterioration in relationships and further isolation, as indicated by one participant's comment: "No one wants to know me. People can't handle my mental health conditions. I feel abandoned."

Table 3: Ability to identify distress as a function of participant group (%)

Are you able to identify the causes of distress?				
	Carers	Autism/ learning disabilities	Dementia	Cognitive impairments
All or most of the time	9	7	0	14
Mostly but not all the time	46	40	60	41
Some of the time	39	47	30	36
Mostly I am unable to identify the causes	7	7	10	9
Never or hardly ever	0	0	0	0

From the perspective of caregivers, the individuals they care for experiencing distress can exhibit anger, dissatisfaction, and may refuse to eat or drink, displaying erratic behavior. Caregivers face challenges in ensuring the physical safety of the person and find it difficult to persuade them to eat and drink. The experienced feelings of distress also affect the work capabilities of the caregiver, as one caregiver mentioned, "As the person gets more and more stressed, I find I cannot care for the other people." Additionally, caregivers highlighted the impact on their own mental well-being, expressing that continuously dealing with or trying to distract from the person's anxiety is mentally and physically exhausting.

A similar pattern emerged when examining the success of strategies to manage distress. The majority of responses indicated that strategies were successful "most of the time" (40-64% of responses) or "sometimes" (36-40% of responses).

Table 4: Success of strategies to cope with distress as a function of participant group (%)

To which extent these strategies are successful?				
	Carers	Autism/ learning disabilities	Dementia	Cognitive impairments
All the time	3	0	10	0
Most of the time	55	60	40	64
Sometimes	39	40	40	36
Almost never	3	0	0	0
Never	0	0	0	0

The open-ended questions asking participants to indicate the causes of distress and the strategies to cope with these feelings provided further insight into the participants' experiences. Based on the responses, the most common causes of distress were physical pain, frustration, environmental factors (such as loud noises or temperature changes), changes in plans, forgetfulness, and lack of communication with family members. Participants mentioned various strategies to cope with distress, with the most common one being avoiding situations that could trigger distress. Other strategies included using distraction methods such as assigning tasks, providing puzzles or quizzes, playing music, better pain management, watching favorite television programs, or simply practicing patience while offering a comforting cup of tea. Patience, in various forms, was frequently mentioned by family caregivers, emphasizing the importance of repetition, explanation, trust, and maintaining the individual's independence within a supportive family environment. These responses indicate several groups of solutions, but the main underlying theme is that when feelings of distress become overwhelming, caregivers and those around the individual need time and patience to try to calm them down.

The first part of the survey revealed the impact of distress on the individuals themselves and the people around them. Interestingly, although the majority of people could recognize some causes of distress, only a small percentage could reliably identify them. Additionally, while the solutions for dealing with distress were reported to be effective most of the time, they were not always successful. This suggests that once the initial stages of distress are not addressed, it becomes challenging to calm the person down and deal with the resulting consequences.

#### 4.2. *SmartSocks as an early indication of feelings of distress*

The final part of the survey was dealing with the participants attitudes towards and expectations of smart sock technology. The following section started with the question of whether participants thought SmartSocks could be useful to them now, in the future, or not at all. Interestingly, 47% autism and learning disabilities participants indicated that they would not find this technology useful, while other populations indicated that they would find this technology useful either now or in the future.

Table 5: Usefulness of SmartSocks as a function of participant group (%)

	Would SmartSocks be useful to you?			
	Carers	Autism/ learning disabilities	Dementia	Cognitive impairments
Yes	41	20	20	27
In the future	43	33	70	36
Not at all	11	47	10	36

The exploration of different functions (Fig 2) per participant group was further investigated. The first step involved examining whether potential covariates, such as the number of distress symptoms and experienced distress severity, correlated with participants' indication of the usefulness of SmartSocks. The Spearman's rho correlation test revealed a significant positive correlation between reported distress severity and SmartSocks usage among the carers' participant group (Spearman's rho = .461,  $p = .015$ ). Similarly, an equivalent analysis showed a strong positive correlation between the number of distress symptoms experienced and the self-reported usefulness of the socks among carers (Spearman's rho = .628,  $p < .001$ ). Interestingly, these correlations between distress severity, number of symptoms, and perceived usefulness of SmartSocks did not significantly correlate among the other participant groups. Since the correlations were significant in at least one participant population, they were included as covariates in the MANOVA analysis. The dependent measures were the usefulness of the different functions ("early indication of pain or anxiety," "identification of sources of distress," "understanding day-to-day levels of distress," and "evidence of distress to provide to health services"), and the independent variables were the functions and participant group (carers, dementia, autism/learning disabilities, and cognitive impairments). The analysis revealed a significant difference in "help to identify sources of distress through alert system" ( $F(3, 89) = 2.99$ ,  $p = .035$ ), and a nearing significance difference in "early indication of pain or anxiety" ( $F(3, 89) = 2.60$ ,  $p = .057$ ). However, the suggested functions of "evidence of distress to provide to health services" and "understanding the overall day-to-day level of distress, anxiety, or pain" did not reach significance ( $p > .05$ ). Post hoc analysis with Bonferroni correction indicated that the differences were at a trend level, where the carers group had higher usefulness scores compared to the cognitive impairments group for both functions, "help to identify sources of distress through

alert system" and "early indication of pain or anxiety" ( $p = .069$  and  $p = .079$ , respectively, Fig. 2).

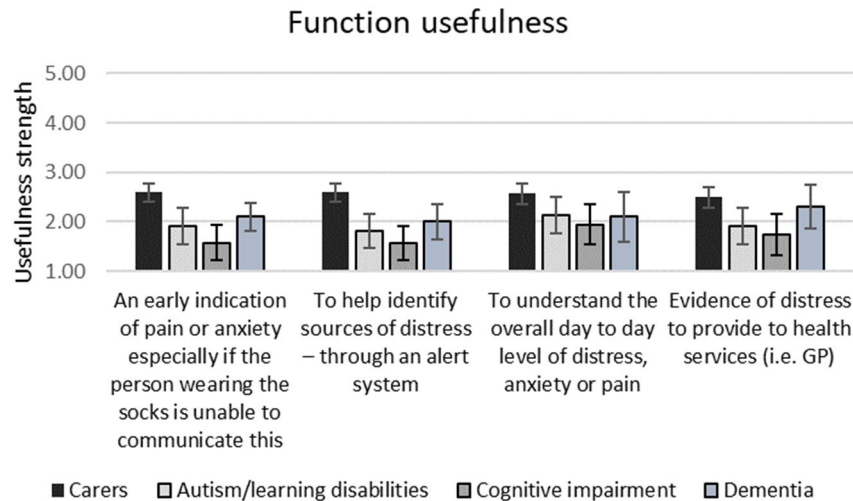


Figure 2: Mean usefulness rating per Smart Sock function distributed among participant groups

## 5. Discussion

The current study aimed to explore the observations and opinions of four participant groups (carers, people living with dementia, autism and learning disabilities, and cognitive impairments) regarding the impacts of experiencing feelings of distress, the noticed causes, and the strategies employed to minimize the negative effects on both the individuals experiencing distress and those around them. Although no significant differences were observed in these aspects, there was a positive correlation between the indication of SmartSocks usefulness in carers' daily lives and the reported severity of distress and number of distress symptoms. In other words, participants who experienced more symptoms and stronger distress were more likely to express a desire to use this technology either now or in the future.

The main objective of the survey was to explore how different populations experience and cope with distress. Interestingly, the responses did not suggest significant differences in distress experiences or coping strategies based on the specific condition. The responses overlapped among the participant groups, indicating a common negative impact of distress regardless of the condition. However, it was unexpected that a majority of participants with autism and learning disabilities expressed a lack of interest in using SmartSocks technology. There are several potential explanations for this finding. Firstly, the relatively low number of participants with autism and learning disabilities meant there were too few responses to provide definitive answers. This needs to be further explored with a larger sample. Secondly, discussions with subject experts revealed that an online survey might not be the most suitable method for capturing the opinions of this population, as only highly functioning and able individuals would be able to focus for approximately 10 minutes to read the text and answer the questions, and these individuals would be less likely to experience significant distress and agitation. Conducting interviews would possibly provide a better method for capturing the opinions and experiences of individuals with autism and learning disabilities. Additionally, an online survey investigating future technology faces the limitation that not all individuals can envision how they would use such technology without relatable examples.

The results also raise ethical questions regarding the perceived usefulness of such technology for the target groups (Tu & Gao, 2021). It is important to note that technology should not be imposed on anyone; rather, it should be perceived as useful and beneficial (Segura Anaya et al., 2018). Two key points emerge from this discussion: (i) the necessity to introduce functionality that provides immediate value for the target population group, whilst also satisfying the needs of clinicians/carers; and (ii) effectively communicating the information and potential benefits in a relatable manner that resonates with the experiences of user groups. These challenges emphasize the importance of future discussions and closer collaboration with ASD and LD participants to gain a better understanding of their needs and challenges, which can then be reflected in the development of SmartSocks.

Despite the limitations, the responses provide evidence and encouragement that SmartSocks technology would be beneficial for paid and family carers, particularly in cases where distress is more prevalent. The survey highlighted the need for examples of how people use the technology and their experiences, which will be further explored through feasibility trials. The survey offers an overview of opinions and directions for the functionality of this technology. The next step involves conducting feasibility trials in several care homes over a two-week period to gather evidence on how the socks can be used and to collect examples and personal experiences from users. These narratives will be discussed in subsequent focus groups, allowing participants from different demographics to explore the potential uses of this wearable technology and provide their feedback.

## **6. Conclusion**

The feelings of distress and agitation have a detrimental impact on individuals living with dementia, autism and learning disabilities, and on their family members and carers. The present study provides evidence regarding how distress is experienced among these participant groups and explores the potential usefulness of wearable technology and its functions for these populations. It is not surprising that the severity of distress and the number of distress symptoms were positively correlated with the intention to use the technology, particularly in the carers group. Moreover, the results suggest that a more personalised approach is necessary to capture the opinions and attitudes of participants with autism and learning disabilities. This study aims to foster a dialogue with the target users of the developing SmartSock technology, facilitating ethical technology development through co-creation between users and technology developers.

## **Acknowledgements**

This research has been supported by the Innovate UK grant (grant number 10045645). The research team would like to thank everybody who was involved and supported this work.

## **References:**

- Amiribesheli, M., & Bouchachia, H. (2018). A tailored smart home for dementia care. *Journal of Ambient Intelligence and Humanized Computing*, 9(6), 1755–1782.
- Aneshensel, C. S., Pearlin, L. I., Levy-Storms, L., & Schuler, R. H. (2000). The Transition From Home to Nursing Home Mortality Among People With Dementia. *The Journals of Gerontology: Series B*, 55(3), S152–S162.
- Bankole, A., Anderson, M., Knight, A., Oh, K., Smith-Jackson, T., Hanson, M. A., Barth, A. T., & Lach, J. (2011). Continuous, non-invasive assessment of agitation in dementia using inertial body sensors. *Proceedings of the 2nd Conference on Wireless Health*, 1–9.

- Broady, T. R., Stoyles, G. J., & Morse, C. (2017). Understanding carers' lived experience of stigma: The voice of families with a child on the autism spectrum. *Health & Social Care in the Community*, 25(1), 224–233.
- Cohen-Mansfield, J. (2001). Nonpharmacologic Interventions for Inappropriate Behaviors in Dementia: A Review, Summary, and Critique. *The American Journal of Geriatric Psychiatry*, 9(4), 361–381.
- Dai, B., Larnyo, E., Tetteh, E. A., Aboagye, A. K., & Musah, A.-A. I. (2020). Factors Affecting Caregivers' Acceptance of the Use of Wearable Devices by Patients With Dementia: An Extension of the Unified Theory of Acceptance and Use of Technology Model. *American Journal of Alzheimer's Disease & Other Dementias*, 35, 153331751988349.
- Dementia Statistics Hub | Alzheimer's Research UK. (n.d.). Dementia Statistics Hub. Retrieved 1 June 2023, from <https://dementiastatistics.org/>
- Goodwin, M. S., Mazefsky, C. A., Ioannidis, S., Erdogmus, D., & Siegel, M. (2019). Predicting aggression to others in youth with autism using a wearable biosensor. *Autism Research*, 12(8), 1286–1296.
- Hedman, E., Miller, L., Schoen, S., Nielsen, D., Goodwin, M., & Picard, R. (2012). Measuring autonomic arousal during therapy. *Proc. of Design and Emotion*, 11–14.
- Herron, R. V., & Wrathall, M. A. (2018). Putting responsive behaviours in place: Examining how formal and informal carers understand the actions of people with dementia. *Social Science & Medicine*, 204, 9–15.
- Home—Skills for Care. (n.d.). Retrieved 1 June 2023, from <https://www.skillsforcare.org.uk/Home.aspx>
- <https://www.kingsfund.org.uk/node/4311>. (2019, December 6). Health and social care funding. The King's Fund. <https://www.kingsfund.org.uk/blog/2019/12/health-and-social-care-funding-who-offers-most>
- Kang, S. J., Choi, S. H., Lee, B. H., Jeong, Y., Hahm, D. S., Han, I. W., Cummings, J. L., & Na, D. L. (2004). Caregiver-Administered Neuropsychiatric Inventory (CGA-NPI). *Journal of Geriatric Psychiatry and Neurology*, 17(1), 32–35.
- Kojima, G., Liljas, A., & Iliffe, S. (2019). Frailty syndrome: Implications and challenges for health care policy. *Risk Management and Healthcare Policy*, Volume 12, 23–30.
- Kolanowski, A., Fick, D. M., Campbell, J., Litaker, M., & Boustani, M. (2009). A Preliminary Study of Anticholinergic Burden and Relationship to a Quality of Life Indicator, Engagement in Activities, in Nursing Home Residents With Dementia. *Journal of the American Medical Directors Association*, 10(4), 252–257.
- Koumpouros, Y., & Kafazis, T. (2019). Wearables and mobile technologies in Autism Spectrum Disorder interventions: A systematic literature review. *Research in Autism Spectrum Disorders*, 66, 101405.
- Kovach, C. R., Noonan, P. E., Schlidt, A. M., & Wells, T. (2005). A Model of Consequences of Need-Driven, Dementia-Compromised Behavior. *Journal of Nursing Scholarship*, 37(2), 134–140.
- Lai, C. (2014). The merits and problems of Neuropsychiatric Inventory as an assessment tool in people with dementia and other neurological disorders. *Clinical Interventions in Aging*.
- Molyneux, G. J., McCarthy, G. M., McEniff, S., Cryan, M., & Conroy, R. M. (2008). Prevalence and predictors of carer burden and depression in carers of patients referred to an old age psychiatric service. *International Psychogeriatrics*, 20(6), 1193–1202.
- Morris, S., Patel, N., Baio, G., Kelly, L., Lewis-Holmes, E., Omar, R. Z., Katona, C., Cooper, C., & Livingston, G. (2015). Monetary costs of agitation in older adults with Alzheimer's disease in the UK: Prospective cohort study. *BMJ Open*, 5(3), e007382–e007382.
- Moyle, W. (2019). The promise of technology in the future of dementia care. *Nature Reviews Neurology*, 15(6), Article 6.
- Moyle, W., Murfield, J., & Lion, K. (2021). The effectiveness of smart home technologies to support the health outcomes of community-dwelling older adults living with dementia: A scoping review. *International Journal of Medical Informatics*, 153, 104513.

- O'Sullivan, G., Whelan, B., Gallagher, N., Doyle, P., Smyth, S., Murphy, K., Dröes, R.-M., Devane, D., & Casey, D. (2023). Challenges of using a Fitbit smart wearable among people with dementia. *International Journal of Geriatric Psychiatry*, 38(3), e5898.
- Ringman, J. M., & Schneider, L. (2019). Treatment Options for Agitation in Dementia. *Current Treatment Options in Neurology*, 21(7), 30. <https://doi.org/10.1007/s11940-019-0572-3>
- Schnaider Beerli, M., Werner, P., Davidson, M., & Noy, S. (2002). The cost of behavioral and psychological symptoms of dementia (BPSD) in community dwelling Alzheimer's disease patients. *International Journal of Geriatric Psychiatry*, 17(5), 403–408.
- Schulz, R., Belle, S. H., Czaja, S., McGinnis, K. A., Stephens, A., & Zhang, S. (2004). Long-term Care Placement of Dementia Patients and Caregiver Health and Well-being. *JAMA*, 292(8), 961.
- Segura Anaya, L. H., Alsadoon, A., Costadopoulos, N., & Prasad, P. W. C. (2018). Ethical Implications of User Perceptions of Wearable Devices. *Science and Engineering Ethics*, 24(1), 1–28.
- Serrano-Aguilar, P. G., Lopez-Bastida, J., & Yanes-Lopez, V. (2006). Impact on Health-Related Quality of Life and Perceived Burden of Informal Caregivers of Individuals with Alzheimer's Disease. *Neuroepidemiology*, 27(3), 136–142.
- Sharmin, M., Hossain, M. M., Saha, A., Das, M., Maxwell, M., & Ahmed, S. (2018). From Research to Practice: Informing the Design of Autism Support Smart Technology. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–16.
- Tsai, C.-F., Hwang, W.-S., Lee, J.-J., Wang, W.-F., Huang, L.-C., Huang, L.-K., Lee, W.-J., Sung, P.-S., Liu, Y.-C., Hsu, C.-C., & Fuh, J.-L. (2021). Predictors of caregiver burden in aged caregivers of demented older patients. *BMC Geriatrics*, 21(1), 59.
- Tu, J., & Gao, W. (2021). Ethical Considerations of Wearable Technologies in Human Research. *Advanced Healthcare Materials*, 10(17), 2100127.
- Weir, E., Allison, C., & Baron-Cohen, S. (2022). Autistic adults have poorer quality healthcare and worse health based on self-report data. *Molecular Autism*, 13(1), 23.





**SECTION-3**  
**HUMAN-ROBOT COLLABORATION**



## **TOWARDS AN ETHICAL FRAMEWORK FOR HUMAN-ROBOT COLLABORATION IN MANUFACTURING: METHODOLOGICAL CONSIDERATIONS**

TIZIANA C. CALLARI, ELLA-MAE HUBBARD, and NIELS LOHSE  
*Intelligent Automation Centre, Wolfson School of Mechanical, Electrical and Manufacturing  
Engineering, Loughborough University, Loughborough*  
E-mail: [T.Callari@lboro.ac.uk](mailto:T.Callari@lboro.ac.uk) [E.Hubbard@lboro.ac.uk](mailto:E.Hubbard@lboro.ac.uk), [N.Lohse@lboro.ac.uk](mailto:N.Lohse@lboro.ac.uk)  
[www. https://www.lboro.ac.uk/research/intelligent-automation/](https://www.lboro.ac.uk/research/intelligent-automation/)

This paper presents the methodological rationale and design of a qualitative scenario-driven Delphi study. Its objective was to gather insights from subject-matter experts in the fields of ethics in robotics, Artificial Intelligence/Machine Learning, (intelligent) technologies in relation to the ethical considerations of the human-robot collaboration (HRC) in the manufacturing industry. While ethical concerns related to HRC, such as the 'displacement/replacement' of workers, 'safety,' 'design and usability,' and 'privacy and data protection,' have been addressed to date, other specific ethical concerns, like supporting the psychological wellbeing of workers, remain under-explored and require further attention. This paper introduces the methodological research process used to derive a list of principles that will form the ethical framework for HRC in manufacturing. Three rounds of data collection were employed to achieve this. In the first questionnaire, experts were asked to provide insights and concerns on emerging ethical challenges, focusing on the HRC ethical gaps using prospective scenarios. In the second questionnaire, a list of candidate principles identified through a data-driven analysis approach was submitted to the experts for review based on the criteria of 'completeness' and 'importance'. Finally, the revised list was ranked to determine the principles that should be included in the consolidated Ethical Framework for HRC in manufacturing.

### **1. Introduction**

#### **1.1. Background information**

Industrial robots and operators have operated separately on production lines for years due to safety concerns. However, economic and environmental factors are now driving the need for a new labour organisation involving increased collaboration between humans and robots to achieve defined objectives [1]. This collaboration, known as human-robot collaboration (HRC), involves the sharing of tasks (on the same component) and the physical workspace (at the same time). The robots are equipped with machine learning (ML) algorithms to enable adaptability to various tasks. To ensure the safety of the human workforce and minimise the risks of accidents caused by collisions with industrial machines, the robots used in manufacturing are designed to be lightweight. Additionally, the shop floor is equipped with interconnected cyber-physical systems, sensors, and cameras that detect and monitor human-robot actions and overall performance.

The installation of collaborative robots, also known as "cobots," aims to enhance the efficiency and quality of industrial production. By taking over repetitive, heavy, and hazardous tasks, cobots allow the human workforce to focus on more advanced cognitive and supervisory roles [2, 3]. However, these changes are challenging how the production shop floor will be designed and implemented, and how the employed human resources will be affected by this physical, technological and social transformation. Therefore, it is paramount to address the ethical and legal implications of HRC in the manufacturing industry.

Issues relating to the ethical implications derived from the synergic collaboration of humans and industrial robots have mainly regarded topics such as 'displacement/replacement' of

workers, 'safety', 'design and usability', 'privacy and data protection' [4]. Indeed, the most debated ethical concern is whether this new industrial revolution (often referred to as Industry 4.0) will create a jobless society [5]. Both at scientific and societal levels, articles present rates of the extent to which the installation of robots is directly proportional to the displacement of workers. A recent study by Oxford Economics [6] estimated that by 2030 approximately 15% of the worldwide workforce might be negatively affected by robotics, as it will automate many tasks currently performed by the human workforce. Although dependent on the different jurisdictions, ethical and legal considerations in HRC encompass a range of topics, such as the collection, use and protection of sensitive data gathered on the manufacturing shop floor. Indeed, workers must be aware of the purpose for which data – both personal and performance – are collected and processed. Another well-known ethical concern in the HRC domain in industrial settings regards workers' physical wellbeing and safety. ISO standards [7-9] and human factors (HF) research face the challenge of advancing theoretical and methodological models capable of integrating and predicting the complex safety-related relationship and collaboration between humans and robots on the factory floor. This specifically pertains to safety requirements in the design of collaborative workstations with cobots, aiming to prevent any physical harm to the human workforce during HRC. The psychological harm that may result from the interaction and task-sharing with robots is an area that has been insufficiently researched [10]. Furthermore, the design and usability of robots are crucial for establishing a safe and responsive interaction with industrial machines on the production line. This raises questions regarding responsibility and liability in situations where robots cause harm to humans during collaboration or when damages occur to production components during HRC. Trust plays a critical role in establishing an effective and dependable HRC, as well as in clarifying human behaviour and intentions in the event of incidents during HRC [11, 12].

In addition to the ethical considerations mentioned above, there are other emerging specific concerns that necessitate further research. These considerations were incorporated into the research design with the objective of developing a comprehensive list of principles that outline the Ethical Framework for HRC in manufacturing.

## **1.2. *Research aim and study objective***

This study forms part of the ISCF Made Smarter Innovation – Smart Cobotics Centre (MSI-SCC) project ([smartcobotics.org.uk](http://smartcobotics.org.uk)), whose aim is to advance smart manufacturing by eliminating barriers and accelerating the widespread use of smart collaborative robotics technology in the UK. As part of this project, Priority Area (PA) 4 aims to define a pathway towards enhanced social acceptability by examining the transformative societal and cultural impact of smart, collaborative automation. The insights gained from this investigation will inform the development of policies and educational strategies that promote long-term sustainability and growth. This research aligns with the overarching objective of PA4.

The overarching aim of this research is to acquire new knowledge regarding the ethical implications of HRC in the manufacturing industry. To accomplish this, subject-matter experts were invited to comment and provide insights into several pressing ethical concerns surrounding HRC in industrial sectors. These concerns include the 'displacement/replacement' of workers, 'privacy and data protection', 'safety', and 'design and usability'.

The objective of this paper is to present the scenario-driven qualitative Delphi approach used to derive a list of principles that will form the ethical framework for HRC in manufacturing. The participants were subject-matter experts in the fields of ethics in robotics, Artificial Intelligence/Machine Learning, and (intelligent) technologies. The research design procedure

involved three rounds of online data collection through questionnaires. In the first questionnaire, experts were asked to share their insights and concerns regarding emerging ethical challenges, with a specific focus on the ethical gaps in HRC, using prospective scenarios. In the second questionnaire, a list of candidate principles, identified through a data-driven analysis approach of the responses from the first questionnaire, was submitted to the experts for review based on the criteria of *completeness* and *importance*. Finally, the revised list was ranked to determine the final principles that should be included in the consolidated Ethical Framework for HRC in manufacturing.

## **2. Overlooked ethical considerations**

This focussed review of the literature served to highlight the gaps that needed to be investigated and reviewed by the Delphi experts.

### **2.1. *Worker (re)skilling to tackle a future jobless society***

The increased adoption of robots across all sectors - including service, military, and manufacturing - raises questions about the future of work and the skills required by workers. Indeed, automation and robotics will perform many routine and non-routine tasks, leading to job displacement, particularly in sectors heavily reliant on manual labour. This presents a significant challenge for workers to retain their jobs in the context of Industry 4.0 [2, 3]. While the literature advocates for worker re-skilling and up-skilling, specific ethical concerns emerge in relation to the abilities and competencies expected of workers to retain their jobs. Creative and problem-solving capabilities are increasingly valued over manual and operational skills

However, ethical considerations arise as to whether employees in future industrial robotics factories should be required to engage in more critical thinking activities to retain their jobs, rather than simply performing routine and non-creative tasks [13, 14]. Furthermore, there will be a demand for new technical and non-technical skills and competencies, necessitating educational institutions to make efforts in designing and/or updating university curricula and professional training programs [15].

### **2.2. *Performance data recording and storage of the human-robot interaction and collaboration***

Manufacturing industries implementing HRC will extensively utilise perception systems, such as cameras and smart sensors, to document the interaction between humans and robots on the shop floor. This integration aims to create a safe and reliable working environment by enabling continuous perception, detection, and monitoring of human-robot interaction (HRI) throughout the production process [16, 17].

Ethical concerns related to data sharing, confidentiality, and privacy, as well as performance data monitoring, are crucial in HRC within manufacturing settings [10, 13]. In the context of Industry 4.0, there is an ongoing exchange of data captured from sensors and robotics involved in HRI on the factory floor, which is then transferred to middle and top management [16, 17]. Understanding how this data, encompassing individual and collaborative human and robot performance, is managed is still an area that requires further exploration. It is crucial to investigate the ethical implications to ensure that the collection of data adheres not only to privacy and data protection protocols but also serves specific safety and performance monitoring purposes. The uncertainty surrounding the ethical use of their data when interacting with robots may lead workers to behave differently [18].

Further research and policy guidelines are necessary within manufacturing and industrial settings to address the ethical implications of performance data monitoring in HRC. These efforts will help elucidate the trade-off between performance, quality, and wellbeing resulting from the integration of HRC [4]. The ethical implications of performance data monitoring warrant additional research and policy guidelines within manufacturing/industrial settings to help clarify the extent of the trade-off between performance, quality and wellbeing as a result of HRC [4].

### **2.3. *Worker wellbeing***

The health and safety of workers during close interactions with robots must be ensured, addressing both the physical and psychological wellbeing of Industry 4.0 workers. In terms of physical aspects, various technologies have been developed to guarantee safe operations, including real-time collision avoidance systems [19]. On the other hand, the psychological wellbeing of workers involves their emotions and perceptions within the human-robot relationship, with trust being identified as a fundamental factor [10, 20]. Additionally, the psychological wellbeing of workers can be influenced by the specific design of task allocation in HRC.

Traditional function allocation and work studies guide industrial human-robot task allocation, which may lack flexibility and potentially lead to psychological distress for workers. This can result in issues like burnout and anxiety when interacting with cobots [13, 21]. Such situations may arise when workers feel pushed by the cobot to achieve higher/faster performance or when operators become frustrated if the robot's actions are perceived as too simple or slow.

While the engineering design of HRC considers physical ergonomics, the mental and psychological impacts of close collaboration on workers are often overlooked and may require further research [10]. In line with this, recent studies aim to understand the role of emotions and embodied experiences in the subjective perception of HRC [13].

### **2.4. *Design and accountability***

In the future, industrial adaptive robots will be trained by AI and machine learning systems, using performance data gathered either from the manufacturing shop floor or directly from skilled human operators. However, during the production process, these trained robots may inadvertently cause disruptions or damage to product components or work tools, resulting in economic consequences such as the need to replace the damaged parts or delays in product delivery. These disruptions affect the technical system and processes and interrupt workers' tasks, schedules, and working conditions, potentially impacting their wellbeing. In the worst-case scenarios, such incidents could lead to accidents that directly jeopardise the safety of human operators [22].

The question arises as to who should be held accountable when these disruptions occur. It is crucial to establish clear legal frameworks to address liability issues and ensure that the affected parties have the means to seek compensation. By addressing this issue, we can also gain a better understanding of human behaviours, intentions, and trust in collaborating with robots [7, 8, 10].

### **3. Methodological rationale**

#### **3.1. *Qualitative Delphi approach***

The Delphi method was originally developed in the late 1940s by researchers at the RAND Corporation during the Cold War. It was initially used to estimate Soviet bombing requirements on potential US industrial targets [23]. Over time, the method has been commonly employed with the goal of reaching a consensus in various areas such as social policy, health practice, and organisational decision-making.

Traditionally, the Delphi method utilises quantitative and/or mixed methods research designs to analyse subjective judgments from domain experts [24, 25]. The method is characterised by several key aspects: (1) gathering multiple perspectives and feedback from experts in the field, (2) managing multiple stages of asynchronous assessment and reassessment, (3) selecting information that has undergone iterative rounds of review by the expert participants, and (4) refining and prioritising the most crucial requirements based on experts' ratings until a consensus is achieved [26, 27]. Typically, Delphi research designs involve three rounds of data collection through questionnaires [28, 29]. Additionally, the Delphi method involves independent individuals (such as the researchers) in each iteration to synthesise and summarise the collected data.

The literature has seen various variations of the traditional Delphi method to the extent that researchers now refer to it as a "research approach" rather than a "method" [30]. Qualitative modifications retain the core characteristics of the traditional Delphi method, but they tend to use open-ended questions instead of structured ones. This allows participants to reflect and comment on emerging themes and collective perspectives derived from their responses [31]. This qualitative process concludes when participants' responses exhibit a convergent trend that addresses the project objective(s) or when sufficient information has been exchanged to achieve information power [32] regarding newly developed themes.

#### **3.2. *Scenario-based approach to elicit expert insights***

Scenario-based approaches are argued to provide the rhetorical setting in which professionals draw from their professional expertise to formulate potential solutions to their needs. Indeed, they have been widely used to inform about human activity (i.e., actors, problems and potential real-life situations) [33-36], and envision future work activity [37-39]. By means of scenarios, experts can broaden perspectives, raise questions and explore the different outcomes associated with a potential alternative of the future.

Prospective scenarios are a technique used in strategic planning to explore and envision potential future outcomes and situations to show how they may develop over time [40]. They are hypothetical narratives or descriptions of possible future states based on a set of assumptions (often derived from a review of up-to-date literature). The goal of prospective scenarios is to provide a range of plausible narratives that challenge assumptions and broaden thinking about potential futures. For this reason, they are used to support creative problem-solving in innovation outputs by involving sequential stages of problem understanding, idea generation, and planning for action [38].

## 4. Research design

### 4.1. Participants

Participants in this study were subject-matter experts in the field of ethics in robotics, AI/ML, (intelligent) technologies. They have been selected based on their scientific reputation and work in ethics.

Overall, 33 experts were recruited. The majority of these experts have backgrounds in legal and ethics subjects, with a specialisation in ethics of AI (17), ethics in Robotics/Roboethics (16), law and ethics of Technology (6), and ethics of Automation (4). Regarding expertise level, 18 experts have more than ten years of experience in the field, while 10 participants have been involved in the topic for 5-10 years (Figure 1). As for their current professional roles and fields of application, they mainly hold positions in academia, as illustrated in the chart below (Figure 2).

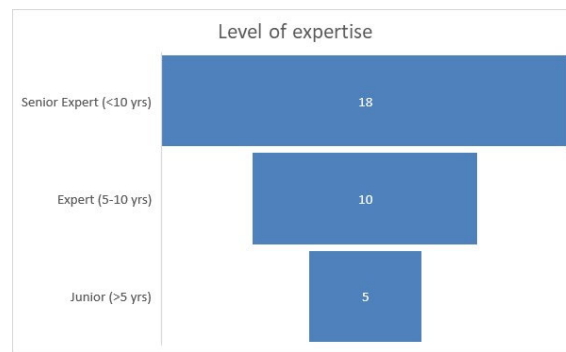


Figure 1: Level of expertise of the study participants

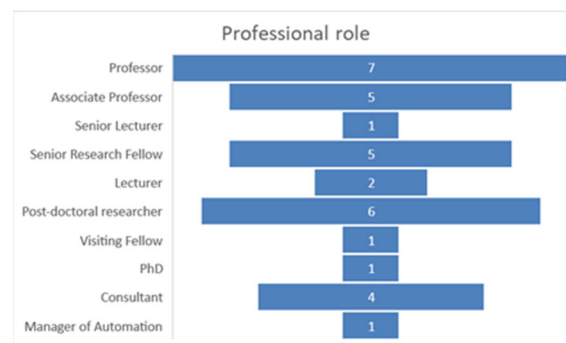


Figure 2: Professional role covered by the study participants

### 4.2. Procedure

Three rounds of data collection were employed, as detailed in what follows.

#### 4.2.1. *Round 1: Experts' divergent thinking on ethical issues in HRC prospective scenarios*

The first questionnaire was launched in February 2023, and the data collection is ongoing (until the end of March). It included four sections: (1) an introduction summarising the study objective; (2) an informed consent section to agree to participate in the research; (3) a participant



form to record primary demographic data and the expert's experience in the Ethics domain; and (4) a final section listing four future-based scenarios to which the recruited experts were asked to comment about what ethical themes and/or considerations the scenarios pose. The four future-based scenarios involved the following:\*

- **SCENARIO (1). Worker (re)skilling.** This scenario aimed to capture the ethical issues regarding the skilling and re-skilling of human workers in the context of human-robot collaboration in the workplace.
- **SCENARIO (2). (Performance) Data monitoring.** This scenario aimed to describe a future-like shop floor where perception systems like cameras and smart sensors will be extensively integrated to ensure a safe and reliable working environment for human-robot collaboration. Information about human and robot performance will be collected and processed using different systems to generate 'intelligence' that could be used to improve production planning and operations through machine learning tools.
- **SCENARIO (3). The psychological side of collaborative tasks in human-robot collaboration.** This scenario aimed to suggest the challenge of future dynamic task allocation in human-robot collaboration, where traditional allocation methods may not efficiently adapt to emergent changes, leading to inflexible collaboration that could impact worker mental and psychological wellbeing.
- **SCENARIO (4). Responsibility/accountability.** This scenario aimed to raise concerns of who should be held accountable for incidents occurring during human-robot collaborations on the shop floor.

Each expert independently completed the questionnaire, sharing their opinions, insights, and recommendations. The responses collected from this first questionnaire were analysed thematically using Braun and Clarke's method [41, 42]. To manage the complexity of the empirical data, NVivo (©Lumivero), a computer-assisted data analysis software, was employed. More than 500 codes were generated and organised into clusters based on overlapping patterns (Figure 3). Initially, themes were developed following the inputs from each of the four scenarios.

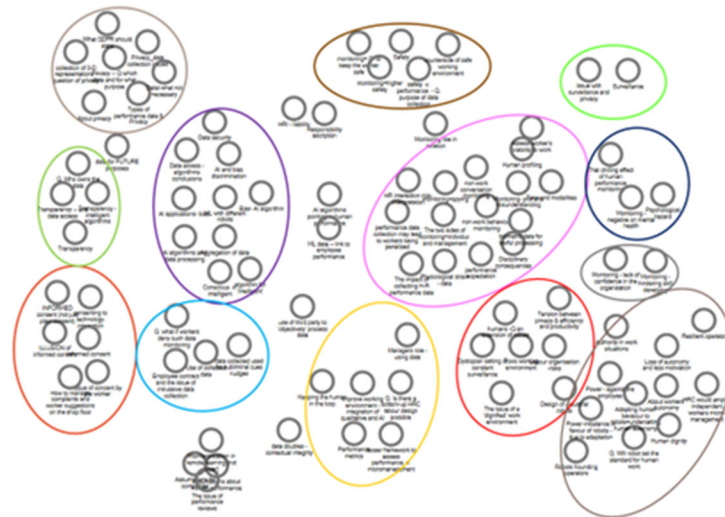


Figure 3: Code clustering and theme development

\*The scenario scripts provide here are a summary of the original one.

The identified data-driven clusters/themes were further reviewed to create a comprehensive two-level codebook using an inductive approach. High-level ethical issues included aspects such as: “Worker Autonomy in Human-Robot Collaboration”, “Worker Agency in Human-Robot Collaboration”, “Monitoring of Human-Robot Collaboration”, “Worker Safety and Physical hazards”, “Worker Wellbeing and Psychological hazards”, etc. Each high-level theme included “attributes” – i.e., sub-level themes specifying the high-level theme. The coded material was analysed at a descriptive level, and this analytical strategy facilitated the identification of a set of potential/candidate principles for each central theme, as provided in Table 1 below. The list of the identified candidate principles formed the basis for the development of the second questionnaire.

Table 1: Preliminary results from Questionnaire 1

High-level theme	Attribute	Candidate principle
Worker Wellbeing and Psychological hazards	Less human-human interaction	Employers should ensure that reducing human-human collaboration on the shop floor and replacing it with human-robot collaboration does not create psychological hazards.

#### 4.2.2. *Round 2: Experts’ detailed feedback on initial provisional principles*

In the second questionnaire, which was circulated in May/June 2023, the experts were given a list of candidate principles for each high-level theme. The objective was to assess each principle based on two criteria: *completeness* and *importance*. To determine the level of *completeness*, the experts were requested to evaluate the accuracy of the outlined description for each principle, and suggest alternative phrasing. The *importance* was rated using a 4-point Likert scale (Irrelevant; Moderately important; Very important; Essential).

The responses will be analysed qualitatively and quantitatively to provide a refined list of principles to suggest in the final round of data collection.

#### 4.2.3. *Round 3: Agreement on a final list of principles*

When this paper was written (June 2023), round three was not circulated yet. However, in the final questionnaire, we plan to ask our participants to select only those revised principles that they believe are fundamental to form the Ethical Framework for HRC in the manufacturing environment. This is meant to evaluate the convergence (i.e., consensus) over a defined list of principles, with a rating for each principle over 90% of agreement.

## 5. Conclusion

This paper presents the rationale for the methodological design of a study, whose overarching aim is to add knowledge to the ethical implications in the field of human-robot collaboration (HRC) in the processing manufacturing industry. Consulting a range of subject-matter experts over a series of rounds of data collection allows for in-depth consideration of potential issues, with diverse and relevant perspectives. Using a set of scenarios based on current research helped to provide appropriate context. Outputs from such work can be validated with relevant stakeholders and used alongside complementary methods to guide future system design.

At the time of writing the paper in June 2023, the data collection phase of the study was still in progress; therefore, the research results were yet to be available for publication. However,

in the methodology section, we presented how the third questionnaire will be developed and how the results will be interpreted.

### Acknowledgements

The present work is supported by EPSRC (Engineering and Physical Sciences Research Council) and ISCF (Industry Strategy Challenge Fund) under the Made Smarter scheme No EP/V062158/1 Made Smarter Innovation - Research Centre for Smart Collaborative Industrial Robotics.

### References

1. Zacharaki, A., et al., *Safety bounds in human robot interaction: A survey*. Safety Science, 2020. **127**: p. 104667.
2. Murphy, R.R., *Co-bots get the good jobs while workers get a human-robot interaction nightmare*. Science Robotics, 2022. **7**(65).
3. Moniz, A.B. and B.-J. Krings *Robots Working with Humans or Humans Working with Robots? Searching for Social Dimensions in New Human-Robot Interaction in Industry*. Societies, 2016. **6**, DOI: 10.3390/soc6030023.
4. Callari, T.C., et al. *Where are we at? A review of the state-of-the-art of the ethical considerations in human-robot collaboration*. in *AHFE 2023 International Conference, (Applied Human Factors and Ergonomics), July 20-24, 2023*. 2023. San Francisco, CA, USA.
5. Veruggio, G., F. Operto, and G. Bekey, *Roboethics: Social and Ethical Implications*, in *Springer Handbook of Robotics*, B. Siciliano and O. Khatib, Editors. 2016, Springer International Publishing: Cham. p. 2135-2160.
6. Oxford Economics, *How robots change the world. Ahat automation really means for jobs and productivity*. 2019, Oxford Economics: Oxford.
7. ISO, *10218-1: Robots and robotic devices — Safety requirements for industrial robots — Part 1: Robots*. 2011.
8. ISO, *10218-2: Robots and robotic devices — Safety requirements for industrial robots — Part 2: Robot systems and integration*. 2011.
9. ISO/TS, *15066: Robots and robotic devices — Collaborative robots*. 2016.
10. Fletcher, S.R. and P. Webb, *Industrial Robot Ethics: The Challenges of Closer Human Collaboration in Future Manufacturing Systems*, in *A World with Robots: International Conference on Robot Ethics: ICRE 2015*, M.I. Aldinhas Ferreira, et al., Editors. 2017, Springer International Publishing: Cham. p. 159-169.
11. Kadar, E.E., A. Köszeghy, and G.S. Virk, *Safety and ethical concerns in mixed human-robot control of vehicles*, in *A World with Robots: International Conference on Robot Ethics: ICRE 2015*, G.S. Virk, et al., Editors. 2017, Springer. p. 135-144.
12. Charalambous, G. and S.R. Fletcher, *A review of a human factors integration approach for the introduction of collaborative industrial robots in manufacturing settings*. Artificial Intelligence, Robots and Ethics, 2019: p. 107-114.
13. van Wynsberghe, A., M. Ley, and S. Roeser, *Ethical Aspects of Human-Robot Collaboration in Industrial Work Settings*, in *The 21st Century Industrial Robot: When Tools Become Collaborators*, M.I. Aldinhas Ferreira and S.R. Fletcher, Editors. 2022, Springer International Publishing: Cham. p. 255-266.
14. Holford, W.D., *The future of human creative knowledge work within the digital economy*. Futures, 2019. **105**: p. 143-154.
15. WEF, *The Future of Jobs Report*. 2020, World Economic Forum: Geneva, CH.
16. Castro, A., F. Silva, and V. Santos *Trends of Human-Robot Collaboration in Industry Contexts: Handover, Learning, and Metrics*. Sensors, 2021. **21**, DOI: 10.3390/s21124113.

17. Bonci, A., et al. *Human-Robot Perception in Industrial Environments: A Survey*. Sensors, 2021. **21**, DOI: 10.3390/s21051571.
18. Winfield, A., *Ethical standards in robotics and AI*. Nature Electronics, 2019. **2**(2): p. 46-48.
19. Kot, R. *Review of Collision Avoidance and Path Planning Algorithms Used in Autonomous Underwater Vehicles*. Electronics, 2022. **11**, DOI: 10.3390/electronics11152301.
20. Charalambous, G., S.R. Fletcher, and P. Webb, *The development of a scale to evaluate trust in industrial human-robot collaboration*. International Journal of Social Robotics, 2016. **8**: p. 193-209.
21. Yam, K.C., et al., *The rise of robots increases job insecurity and maladaptive workplace behaviors: Multimethod evidence*. Journal of Applied Psychology, 2023. **108**: p. 850-870.
22. Heinzmann, J. and A. Zelinsky, *A Safe-Control Paradigm for Human-Robot Interaction*. Journal of Intelligent and Robotic Systems, 1999. **25**(4): p. 295-310.
23. Dalkey, N. and O. Helmer, *An experimental application of the Delphi method to the use of experts*. Management Science: Journal of the Institute of Management Sciences, 1963. **9**(3): p. 458-67.
24. Mead, D. and L. Moseley, *The use of the Delphi as a research approach*. Nurse Researcher, 2001. **8**(4): p. 4-23.
25. Tapio, P., et al., *The unholy marriage? Integrating qualitative and quantitative information in Delphi processes*. Technological Forecasting and Social Change, 2011. **78**(9): p. 1616-1628.
26. Zartha Sossa, J.W., W. Halal, and R. Hernandez Zarta, *Delphi method: analysis of rounds, stakeholder and statistical indicators*. Foresight, 2019. **21**(5): p. 525-544.
27. Hirschhorn, F., *Reflections on the application of the Delphi method: lessons from a case in public transport research*. International Journal of Social Research Methodology, 2019. **22**(3): p. 309-322.
28. Hasson, F., S. Keeney, and H. McKenna, *Research guidelines for the Delphi survey technique*. Journal of Advanced Nursing, 2000. **32**(4): p. 1008-1015.
29. Okoli, C. and S.D. Pawlowski, *The Delphi method as a research tool: an example, design considerations and applications*. Information & Management, 2004. **42**(1): p. 15-29.
30. Brady, S.R., *Utilizing and Adapting the Delphi Method for Use in Qualitative Research*. International Journal of Qualitative Methods, 2015. **14**(5): p. 1609406915621381.
31. Fusch, P. and L.R. Ness, *Are We There Yet? Data Saturation in Qualitative Research*. The Qualitative Report, 2015. **20**: p. 1408-1416.
32. Braun, V. and V. Clarke, *To saturate or not to saturate? Questioning data saturation as a useful concept for thematic analysis and sample-size rationales*. Qualitative research in sport, exercise and health, 2021. **13**(2): p. 201-216.
33. Carroll, J.M., *Introduction to this Special Issue on "Scenario-Based System Development"*. Interacting with Computers, 2000. **13**(1): p. 41-42.
34. Carroll, J.M., *Making use: scenario-based design of human-computer interactions*. 2000, Cambridge, MA: MIT Press.
35. Carroll, J.M., *Five reasons for scenario-based design*. Interacting with Computers, 2000. **13**(1): p. 43-60.
36. Stanton, N.A., *Human factors methods: a practical guide for engineering and design*. 2005, Hampshire Ashgate Publishing Ltd.
37. Sinclair, M., *Ergonomics issues in future systems*. Ergonomics, 2007. **50**(12): p. 1957-1986.
38. Nelson, J., et al., *Generating prospective scenarios of use in innovation projects*. Le travail humain, 2014. **77**(1): p. 21-38.

39. Nelson, J., S. Buisine, and A. Aoussat, *Anticipating the use of future things: Towards a framework for prospective use analysis in innovation design projects*. Applied Ergonomics, 2013. **44**(6): p. 948-956.
40. Oliveira, A.S., et al., *Prospective scenarios: A literature review on the Scopus database*. Futures, 2018. **100**: p. 20-33.
41. Braun, V. and V. Clarke, *Using thematic analysis in psychology*. Qualitative Research in Psychology, 2006. **3**(2): p. 77-101.
42. Braun, V. and V. Clarke, *Thematic analysis: a practical guide*. 2022, London: SAGE Publications Ltd.

## **HOW CAN HUMAN-ROBOT COLLABORATION IMPROVE OPERATORS' WORKING CONDITIONS AND WELLBEING IN AIRCRAFT FUEL TANK INSPECTION: A MIXED-METHODS USER-CENTRED APPROACH**

VAISHNAVI SASHIDHARAN<sup>1</sup>, IVETA EIMONTAITE<sup>1</sup>, SARAH R. FLETCHER<sup>1</sup>, NIKOS DIMITROPOULOS<sup>2</sup>, SOTIRIS MAKRIS<sup>2</sup>, GEORGE MICHALOS<sup>2</sup>, IGAL ISRAELI<sup>3</sup> AND SCOTT TUCKER<sup>3</sup>

<sup>1</sup>*School of Aerospace, Transport and Manufacturing, Cranfield University, College Road, Cranfield, MK43 0AL, United Kingdom*

*E-mail: [V.Sashidharan@cranfield.ac.uk](mailto:V.Sashidharan@cranfield.ac.uk), [Iveta.Eimontaite@cranfield.ac.uk](mailto:Iveta.Eimontaite@cranfield.ac.uk), [s.fletcher@cranfield.ac.uk](mailto:s.fletcher@cranfield.ac.uk), [www.cranfield.ac.uk](http://www.cranfield.ac.uk).*

<sup>2</sup>*Laboratory for Manufacturing Systems and Automation (LMS), Department of Mechanical Engineering and Aeronautics, University of Patras, Patras, 26504, Greece*

*E-mail: [dimitropoulos@lms.mech.upatras.gr](mailto:dimitropoulos@lms.mech.upatras.gr), [makris@lms.mech.upatras.gr](mailto:makris@lms.mech.upatras.gr), [michalos@lms.mech.upatras.gr](mailto:michalos@lms.mech.upatras.gr), [www.lms.mech.upatras.gr](http://www.lms.mech.upatras.gr).*

<sup>3</sup>*Advanced Design & Development, Aviation Group, Israel Aerospace Industries Ltd., Tel Aviv, Israel.*

*E-mail: [igisraeli@iai.co.il](mailto:igisraeli@iai.co.il), [tucker@iai.co.il](mailto:tucker@iai.co.il)*

*[www.iai.co.il](http://www.iai.co.il).*

Human-robot collaboration (HRC) is a means through which Industry 5.0 can achieve its goal of improving working conditions, wherein robots can perform tasks that are unsafe or unhealthy for human operators to perform. Aircraft fuel tank maintenance is an example of such an industrial process where working conditions need to be improved. An aircraft fuel tank is an inherently unsafe environment, and its confined workspace adversely affects the health of the operators who inspect it. The current paper describes how HRC can be utilised for its ability to elevate working conditions and improve operator wellbeing in the aircraft fuel tank maintenance process at Israel Aerospace Industries Ltd. User-centred research methods like observations, interviews and eye-tracking were employed to analyse the physical and psychological challenges faced by the operators, and to capture the aspects of the process that cannot be automated. The findings of this analysis are discussed in consideration of their implications for HRC implementation. Finally, the future steps of the user-centred research plan to ensure successful HRC and operator wellbeing are explained.

### **1. Introduction**

The primary ideology of Industry 5.0 is to bring back the human element to manufacturing industries through effective collaboration between the human workforce and smart machinery (Maddikunta et al., 2022). The implementation of human-robot collaboration (HRC) in several industries (e.g., aerospace, automotive, white goods) is a direct outcome of this ideology. Another result of Industry 5.0's focus on the human element is workforce sustainability, i.e., a recognition of sustaining the physical and psychological wellbeing of industrial operators. A major factor that impacts operator wellbeing is the work environment. Industrial work environments are often unsafe or physically uncomfortable and have long-lasting adverse effects on the operators' psychological and physical health. In Industry 5.0, HRC is being increasingly

used as a means to improve working conditions and thereby improve operator wellbeing. In fact, some researchers argue that the principal aim of human-robot hybrid working production systems should be to improve working conditions (Dimitropoulos et al., 2021).

Safer work environments can be enabled through HRC, wherein the robots perform tasks in environments that would be dangerous for a human operator, e.g., tasks that require the manipulation or movement of high payloads or tasks that are performed in unsafe environments. Robots can also perform tasks that cause fatigue and bodily strain to the operators performing them, e.g., tasks that require a high level of repeatability and accuracy (Charalambous et al., 2015). However, it is important to retain certain tasks for operators to perform, as human abilities of cognitive reasoning, adjustability, and tolerances to variability in the environment make humans an indispensable production resource (Dimitropoulos et al., 2021).

By efficiently utilizing the respective strengths of humans and robots, successful HRC can ensure that human cognitive abilities are maximised whilst safer and comfortable working environments are created.

## **2. HRC in the aerospace industry**

Several applications of HRC in the aerospace industry have resulted in the betterment of operators' working conditions. Examples of HRC implementation in the aerospace industry are the manufacturing of actuation systems (*Aerospace Industry Automates Cnc Tasks with Robotics*, n.d.), positioning and manipulation of large carbon-fibre parts on tooling (Dimitropoulos et al., 2021), and alignment and fastening key assembly components (Pérez et al., 2020). These applications of HRC were carried out for goals of meeting increasing demand, reducing manufacturing costs and time, and reducing the need for operators to perform non-value adding activities. Alongside meeting these desired aims, the HRC applications in these scenarios also resulted in safer working conditions for the operators.

Therefore, the aerospace industry should critically consider HRC implementation in industrial processes that are dangerous for operators' physical and mental health. The inspection and maintenance of an aircraft fuel tank is a prime example of such a process. Aircraft fuel tanks are typically located in the wings. The inspection and maintenance process of these tanks involve several challenges and health risks. The operator's access to the fuel tank is difficult due to limited space available for the operator. This results in the operators assuming non-ergonomic positions to enter the wing, and they must remain in these non-ergonomic postures while inspecting and carrying out maintenance of the tank. The physical stresses of these postures lead to reduced concentration levels and can result in the operators missing out on defects. Moreover, the working environment of the fuel tank is inherently unsafe due to the potential residue of fumes in the tank, which has hazardous effects on the health of the operator (Heilemann et al., 2021). The narrow space within the fuel tank increases the toxicity of these fumes and chemicals, and can lead to explosions even in small amounts (Gaina, 2019). Additionally, as the operator can remain in the fuel tank for a duration of over thirty minutes at a time, the process is time consuming as well.

Research on other confined workspaces has demonstrated that operators' mental health suffers when they work in a physically restrictive environment. Deep underground miners reported elevated mental health symptoms such as somatisation, anxiety (Xie et al., 2020) and claustrophobia (Soh et al., 2016).

Thus, the aircraft fuel tank inspection is a process that has a definite need for an improvement of working conditions and an addressal of operator wellbeing. HRC

implementation in this process can address the safety issues the operators face and improve wellbeing factors, alongside speeding up the process. In consideration of these benefits of HRC, this case has been taken up by “CONVERGING”, a European research project that aims to provide firms with “innovative, adaptable production systems that utilise AI-based cognition, big data, digital twins, and other technologies” ( *CONVERGING Project |European Commission*, 2022.). Amongst the four industrial use-cases the project is focussed on, the fuel tank maintenance process at Israel Aerospace Industry Ltd. (IAI) is a work environment that is being explored for the implementation of HRC. To ensure that the human element is centralized in the design and implementation of HRC, CONVERGING places its primary focus on the end-user of the HRC, i.e., the industrial operators. This approach to understanding and analysing the current industrial process, and designing the new automated process is known as the user-centred research.

### **3. User-centred research**

User-centred research is an iterative process wherein data from the end-user (in this case, the industrial operators) should be collected and analysed in regular loops of the development of the new technology (Pais et al., 2022). This method of simultaneous technological design and data collection from the operators, allows the design process to be developed based on the needs of the operators working with or using the new technology. Consequently, this will enable greater acceptance of the new innovations by the operators.

In keeping with the specific goal of introducing HRC to improve working conditions, the existing industrial process and operators’ experiences should be studied to identify the specific aspects of the industrial tasks that are unsafe and unergonomic for the operators. This will not only inform the design of task-division in HRC but will also highlight operator wellbeing issues that can be rectified through HRC.

The current paper presents this user-centred approach to studying and analysing the task of fuel tank inspection at IAI. The paper will present the protocol adopted to identify the key human factors contributing to the existing process and discuss the findings of the data collected. The paper will then discuss the future steps of the protocol to ensure effective HRC implementation with regard to physiological, psychological, and organizational wellbeing of the operators.

## **4. Method**

### **4.1. Participants**

Participants were two male operators with 12 and 7 years of experience in the process of non-destructive testing (NDT) inspection and visual inspection respectively. One operator’s role in the company was to perform visual inspection, while the second operator’s role was performing quality control. Both operators performed parts of the process that was observed and recorded.

### **4.2. Procedure**

CONVERGING aims to introduce new technologies and equipment for the inspection of the aircraft fuel tanks. A smart collaborative robot will be implemented to inspect and perform several of the maintenance tasks as required, with the use of appropriate vision sensors combined with decision making algorithms to detect damages and identify the characteristics of these damages.



In order to design the HRC system for this process, CONVERGING aims to collect a comprehensive benchmark of human factors in the analysis of IAI's fuel tank inspection processes. A user-engagement plan has been developed for this purpose. This is a two-step plan to obtain data about the effects of the industrial process on the operators, and on the operators' subjective and objective attitudes towards the existing process. The first step has been conducted and data collection and analysis has been completed. The first step will be explained below, and the second step will be discussed in the "Future Directions" section of the paper.

#### 4.2.1. *Step 1*

The first step of the protocol was a physical visit to the factory of the use-case.

In this visit, there are three methods through which data is collected:

- i. The researchers observe the operators carrying out all the tasks involved in the process. (Although, this was limited due to compact space within the wing).
- ii. Short semi-structured interviews with the operators are conducted to understand which aspects of the tasks they enjoy and find difficult, the points at which they are prone to errors and how they fix the errors, and key decision-making points.
- iii. The operators were asked to wear eye-tracking glasses (SMI Eye Tracking Glasses (SensoMotoric Instruments ETG 1.7)) while performing the tasks. The eye-tracking data was analysed using SensoMotoric's BeGaze© eye-tracking analysis software, utilising the Area of Interest (AOI) semantic gaze mapping. The eye-tracking data was then examined in terms of dwell time (%) recorded within predefined AOIs.

The information collected from observations and eye-tracking data, along with the short interviews with operators were analysed together to indicate which aspects of the task require the highest cognitive load, are enjoyable, easy to perform, and can be improved with new technology or design.

## 5. Findings

The interviews revealed that the operator's main dissatisfaction with the task was the prolonged position they had to assume due to the narrow space of the work environment. These positions become increasingly hard to sustain as the operators grow older. One operator indicated that over time, working in these uncomfortable postures has impacted his psychological health. The interviews also revealed that the operators cannot be claustrophobic or inflexible, as they would not be able to perform the task. In fact, during the observation of the NDT testing, 60% of the time was spent on moving within the wing (relocating between different chambers and repositioning) while only 40% of time was spent to perform the inspection.

Task observation and interviews also revealed that the process is reliant on procedural knowledge, but also heavily dependent on tacit knowledge the operators develop through experience. The operators mentioned that certain instances frequently arise where they are required to conduct the inspection using tactile information from the environment. The eye tracker video footage illustrated the operators' comments on the reliance on tactile knowledge. In the video captured, certain instances arose where the operator had to perform blind inspection as the inspection area was partially obstructed, and the operator used tactile comprehension to complete this part of the inspection (figure 1).

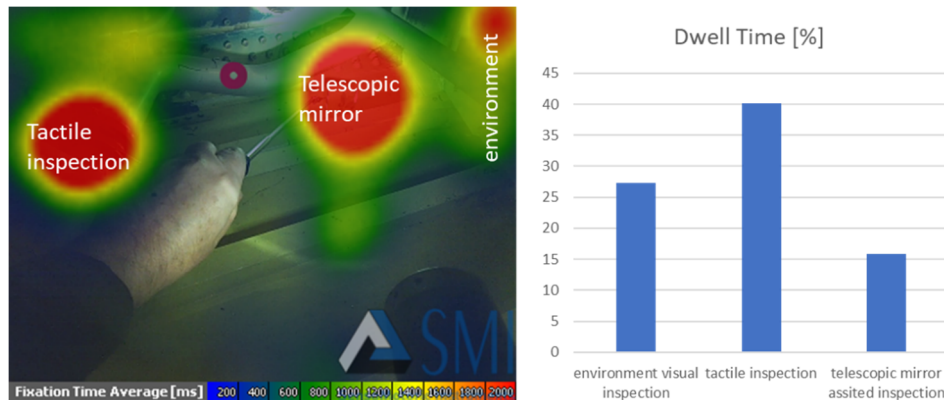


Figure 1: dwell time on the environment visual inspection, telescopic mirror assisted visual inspection and tactile inspection (left - heat map, right – dwell time percentage)

## 6. Discussion and Implications

The user-centred analysis of the task illustrates that the physical discomfort in performing the task has adverse impacts on the operators' psychological and physical wellbeing. Therefore, this suggests that the new HRC system should aim at eliminating the need to spend prolonged times within the wing, as this will yield improvements to the operators' physical and psychological safety. Furthermore, the analysis also demonstrated that the process requires the use of the operators' tacit knowledge and tactile comprehension. This implies that the new HRC system must ensure that the operators are the final decision makers in the inspection process. They need to be the authority on recognizing faults and ensuring that they are fixed.

The HRC system should accommodate these critical human factors of the process and be designed accordingly. This serves as feedback to the engineers and designers for appropriate task-design between the robot and operator. The initial proposed solution is integrating a semi-autonomous collaborative robot (cobot) that will enter the inspection area to detect and remove foreign object debris, as well as perform NDT inspection. However, the entire performance of the cobot will be monitored and guided by the operator, and the data picked up by the cobot's vision sensors will be sent to the operator. If the cobot detects cause for maintenance that it cannot fix, this information is relayed to the operator who can go into the fuel tank and fix the defect.

Hence, the proposed design ensures that the operator's time in the fuel tank is notably reduced. Since the inspection has already been conducted by the cobot, the operator's time in the fuel tank is limited to the time spent on fixing defects, and not on searching and locating the defects. This proposed HRC design also makes sure that the operator can trust the cobot, since they will be notified of the cobot's actions at every moment of the process. The cobot's actions will be predictable and controllable and will not increase the operator's cognitive demand. Importantly, this design allows the operators to be the final decision makers in the process, and apply their tacit knowledge as required.

By reducing the amount of time that the operators spend inside the fuel tank, the implementation of HRC allows an improvement in operators' physical and mental health. The operators would no longer be required to assume uncomfortable positions for large periods of time. This would ensure that their job does not cause body-part discomfort. Improving working conditions and reducing bodily discomfort would consequently improve the **mental workload**

of the operators. Mental workload is a multidimensional construct that is impacted by the characteristics of the task (e.g., demands), operator (e.g., skills, cognitive abilities) and the environment (e.g., physical working conditions) (Young et al., 2015). A very low or very high mental workload could risk the operator's safety and health, and adversely affect the task performance (Rubio et al., 2004). Spending long periods of time in uncomfortable positions and in an unsafe work environment causes undue stress on operators thereby increasing their mental workload higher than the appropriate amount. Additionally, an important factor associated with mental workload is **situational awareness**. This refers to an operator's knowledge of their work surroundings, and comprehension of the current and future operations that need to be carried out (Endsley, 1988). Poor situational awareness is associated with increased errors and accidents (Taylor & Selcon, 1994) and performance failures (Durso & Sethumadhavan, 2008). A physically constraining environment can potentially reduce situational awareness as the operator's physical discomfort can prevent them from accurately comprehending the processes in their task. Therefore, the HRC implementation would also benefit the operators' situational awareness.

Furthermore, data from the user-centered analysis of the task demonstrated that the operators rely on their procedural knowledge and experience. This indicates that the operator must continue to play a key role in the task, even after the HRC implementation. In the new design with the robot, a majority of the operator's task time is now spent on inspection and maintenance rather than entering and positioning themselves within the tank. This would allow a higher engagement with the actual task that the operator was hired and trained for, compared to the current process. Hence with the HRC system, there can be a greater use and application of the operator's skill set and experience. This improved use and valuation of the operator's abilities can improve the operators' attitudes and motivation towards their jobs. Work factors such as **organizational commitment and motivation** and **occupational self-efficacy** could increase with the implementation of HRC. Organizational commitment refers to an employee's identification, level of psychological involvement, loyalty, and sense of belongingness to the organization, (Cook & Wall, 1980) and occupational self-efficacy is an individual's perception of their competence to successfully fulfill their occupational tasks. Research has shown that self-efficacy is positively linked to job satisfaction (Rigotti et al., 2008). With an increase of these occupational factors, the operators' occupational wellbeing would see an uptick with the implementation of HRC.

## 7. Future Directions (Step 2)

While step 1 assessed the operators' immediate physical and psychological concerns and challenges arising from the current process, step 2 will assess these aspects further, as well as measure the organizational and individual factors that are interrelated to the operator's technological engagement. As mentioned in the previous section, the implementation of HRC should ideally decrease operators' bodily discomfort and improve situation awareness and mental workload. Occupational factors like occupational self-efficacy, and organization commitment and motivation should also be positively impacted. Hence, Step 2 involves administering online questionnaires to a statistically larger number of operators to obtain quantitative data on how these factors are impacted. The following is a list of the questionnaires that will be administered:

- 1) Work-related body-part discomfort scale (Cameron, 1996)
- 2) Mental workload NASA TLX (Hart & Staveland, 1988)
- 3) Situational awareness scale (Taylor & Selcon, 1990)

- 4) Organizational commitment and motivational scale (Cook & Wall, 1980)
- 5) Occupational self-efficacy (Rigotti et al., 2008).

The user-centred approach is an iterative process wherein a user-centred analysis of the industrial processes is required before, during and after the implementation of HRC to consistently identify if HRC implementation requires further development or modifications. The above questionnaires will be administered before and after the pilot of the HRC as this will serve as a comparative tool to assess if the HRC implementation has improved the operators' physical comfort, psychological health (situational awareness, mental workload), and job attitudes (occupational commitment and self-efficacy).

In addition to the above questionnaires, a technology readiness level scale (Rose, 2010) will be administered prior to the implementation of the HRC. Technology readiness refers to an individual's likelihood of accepting and using new technology. Measuring the operators' technology readiness level helps understand their propensity to trust, accept and collaborate with the robot. If the readiness level is low, the organization will need to develop strategies to increase their level of readiness. This could be steps such as upskilling the operators, communicating the changes and purpose of the new HRC system to the workforce, demonstrate support and belief in the HRC system, and identify a process champion for new implementations (Charalambous et al., 2015).

It is important to state that future studies investigating the improvement of working conditions in aircraft fuel tanks should aim to avoid the shortcomings of the current study. The task analysis in the current study was conducted on the inspection process of two areas of the wing which were easily accessible for the operators. Following studies should attempt to analyze the inspection conducted on other areas of the wing as well. Furthermore, the current study had two operators as participants, which is a small number. The recruitment of the participants was dependent on convenience and availability of the operators at the time of the site visit. Hopefully, future studies can recruit a larger number of operators to generate data that is more representative of their experience.

## **6. Conclusion**

The current paper presents an example of how HRC can be a solution to improve industrial working conditions and operator wellbeing whilst ensuring ethical implementation of technology within the working environment. The paper demonstrates the user-centred human-factors analysis conducted for the design of an HRC system in aircraft fuel tank maintenance at IAI. The key physical and psychological challenges of the task were identified, as well as the important aspects of the industrial task that are crucial for its successful completion. The implications of this analysis were discussed, in terms of how the human-factors analysis serves as feedback for the design of the HRC. Finally, the next steps in this process of HRC implementation were highlighted.

Hence, the CONVERGING project hopes to show that with a central focus on the industrial operators, strategies of automating industrial processes uphold the human-centric ideologies of Industry 5.0 and can successfully meet other financial and organizational goals.

## **Acknowledgements**

This research has been supported by the EU project "CONVERGING Social industrial collaborative environments integrating AI, Big Data and Robotics for smart manufacturing". This project has received funding from the European Union's Horizon Europe research and

innovation program under grant agreement No 101058521. The authors would like to thank all consortium members for their contribution.

## References

- Aerospace industry automates cnc tasks with robotics*. (n.d.). Retrieved 5 April 2023, from <https://www.universal-robots.com/case-stories/whippany-actuators-systems/>.
- Cameron, J. A. (1996). Assessing work-related body-part discomfort: Current strategies and a behaviorally oriented assessment tool. *International Journal of Industrial Ergonomics*, 18(5), 389–398. [https://doi.org/10.1016/0169-8141\(95\)00101-8](https://doi.org/10.1016/0169-8141(95)00101-8)
- Charalambous, G., Fletcher, S., & Webb, P. (2015). Identifying the key organisational human factors for introducing human-robot collaboration in industry: An exploratory study. *The International Journal of Advanced Manufacturing Technology*, 81(9–12), 2143–2155. <https://doi.org/10.1007/s00170-015-7335-4>
- Cook, J., & Wall, T. (1980). New work attitude measures of trust, organizational commitment and personal need non-fulfillment. *Journal of Occupational Psychology*, 53(1), 39–52. <https://doi.org/10.1111/j.2044-8325.1980.tb00005.x>
- Dimitropoulos, N., Michalos, G., & Makris, S. (2021). An outlook on future hybrid assembly systems—The Sherlock approach. *Procedia CIRP*, 97, 441–446. <https://doi.org/10.1016/j.procir.2020.08.004>
- Durso, F. T., & Sethumadhavan, A. (2008). Situation Awareness: Understanding Dynamic Environments. *Human Factors*, 50(3), 442–448. <https://doi.org/10.1518/001872008X288448>
- Endsley, M. R. (1988). Design and Evaluation for Situation Awareness Enhancement. *Proceedings of the Human Factors Society Annual Meeting*, 32(2), 97–101. <https://doi.org/10.1177/154193128803200221>
- Gaina, M.-G. (2019). Dangerous entry into the aircraft fuel tank – Introduction of mobil robot. *INCAS BULLETIN*, 11(2), 97–110. <https://doi.org/10.13111/2066-8201.2019.11.2.8>
- Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Human mental workload* (pp. 139–183). North-Holland. [https://doi.org/10.1016/S0166-4115\(08\)62386-9](https://doi.org/10.1016/S0166-4115(08)62386-9)
- Heilemann, F., Dadashi, A., & Wicke, K. (2021). Eeloscope—Towards a Novel Endoscopic System Enabling Digital Aircraft Fuel Tank Maintenance. *Aerospace*, 8(5), Article 5. <https://doi.org/10.3390/aerospace8050136>
- Maddikunta, P. K. R., Pham, Q.-V., B, P., Deepa, N., Dev, K., Gadekallu, T. R., Ruby, R., & Liyanage, M. (2022). Industry 5.0: A survey on enabling technologies and potential applications. *Journal of Industrial Information Integration*, 26, 100257. <https://doi.org/10.1016/j.jii.2021.100257>
- Pais, S., Petrova, K., & Parry, D. (2022). Enhancing System Acceptance through User-Centred Design: Integrating Patient Generated Wellness Data. *Sensors*, 22(1), Article 1. <https://doi.org/10.3390/s22010045>
- Pérez, L., Rodríguez-Jiménez, S., Rodríguez, N., Usamentiaga, R., García, D. F., & Wang, L. (2020). Symbiotic human–robot collaborative approach for increased productivity and

- enhanced safety in the aerospace manufacturing industry. *The International Journal of Advanced Manufacturing Technology*, 106(3–4), 851–863. <https://doi.org/10.1007/s00170-019-04638-6>
- Rigotti, T., Schyns, B., & Mohr, G. (2008). A Short Version of the Occupational Self-Efficacy Scale: Structural and Construct Validity Across Five Countries. *Journal of Career Assessment*, 16. <https://doi.org/10.1177/1069072707305763>
- Rose, J. (2010). *Technology readiness and segmentation profile of mature consumers*. [https://core.ac.uk/display/11041044?utm\\_source=pdf&utm\\_medium=banner&utm\\_campaign=pdf-decoration-v1](https://core.ac.uk/display/11041044?utm_source=pdf&utm_medium=banner&utm_campaign=pdf-decoration-v1)
- Rubio, S., Díaz, E., Martín, J., & Puente, J. M. (2004). Evaluation of Subjective Mental Workload: A Comparison of SWAT, NASA-TLX, and Workload Profile Methods. *Applied Psychology: An International Review*, 53(1), 61–86. <https://doi.org/10.1111/j.1464-0597.2004.00161.x>
- Social industrial collaborative environments integrating AI, Big Data and Robotics for smart manufacturing | CONVERGING Project | Fact Sheet | HORIZON | CORDIS | European Commission*. (n.d.). Retrieved 11 April 2023, from <https://cordis.europa.eu/project/id/101058521>.
- Soh, C.-K., Christopoulos, G., Roberts, A., & Lee, E.-H. (2016). Human-centered Development of Underground work Spaces. *Procedia Engineering*, 165, 242–250. <https://doi.org/10.1016/j.proeng.2016.11.796>
- Taylor, R. M., & Selcon, S. J. (1990). Cognitive Quality and Situational Awareness with Advanced Aircraft Attitude Displays. *Proceedings of the Human Factors Society Annual Meeting*, 34(1), 26–30. <https://doi.org/10.1177/154193129003400107>
- Taylor, R. M., & Selcon, S. J. (1994). Situation in Mind: Theory, Application and Measurement of Situational Awareness. *Situational Awareness in Complex Systems*, 69–79.
- Xie, H., Liu, J., Gao, M., Liu, Y., Ma, T., Lu, Y., Li, C., Wang, M., Zhang, R., Wu, J., Zou, J., Liu, S., & Li, W. (2020). Physical symptoms and mental health status in deep underground miners: A cross-sectional study. *Medicine*, 99(9), e19294. <https://doi.org/10.1097/MD.00000000000019294>
- Young, M. S., Brookhuis, K. A., Wickens, C. D., & Hancock, P. A. (2015). State of science: Mental workload in ergonomics. *Ergonomics*, 58(1), 1–17. <https://doi.org/10.1080/00140139.2014.956151>

## TO COLLABORATE OR NOT TO COLLABORATE? HOW TO DETERMINE THE MOST SUITABLE LEVEL OF AUTOMATION TO INCREASE WORKFORCE SUSTAINABILITY AND PRODUCTION EFFICIENCY

SACHA GODHANIA<sup>1</sup>, IVETA EIMONTAITE<sup>1</sup> AND SARAH R. FLETCHER<sup>1</sup>, ANA GONZALEZ  
SEGURA<sup>2</sup>, ELOY RUIZ<sup>3</sup>, ANA PAOLA CARO OSPINA<sup>3</sup>

<sup>1</sup>*Cranfield University, College Rd, Cranfield,  
Wharley End, Bedford, MK43 0AL, United Kingdom*

*E-mail: [Sacha.Godhanian@cranfield.ac.uk](mailto:Sacha.Godhanian@cranfield.ac.uk), [Iveta.Eimontaite@cranfield.ac.uk](mailto:Iveta.Eimontaite@cranfield.ac.uk), [s.fletcher@cranfield.ac.uk](mailto:s.fletcher@cranfield.ac.uk)  
[WWW.CRANFIELD.AC.UK](http://WWW.CRANFIELD.AC.UK)*

<sup>2</sup>*NTT DATA, Avda. Cortes Valencianas 39, 9c, 46015,  
Valencia, Spain*

*E-mail: [ana.gonzalez.segura@nttdata.com](mailto:ana.gonzalez.segura@nttdata.com)  
[nttdata.com](http://nttdata.com)*

<sup>3</sup>*Carrer Mas del Conde, 5. Polígono Industrial.,  
Riba-roja de Tàrrida (Valencia), Spain*

*E-mail: [e.ruiz@andreuworld.com](mailto:e.ruiz@andreuworld.com), [ap.carro@andreuest.com](mailto:ap.carro@andreuest.com)  
[andreuworld.com](http://andreuworld.com)*

With the beginning of Industry 5.0, semi-automated processes such as collaborative robots have become increasingly popular. However, currently, there is a lack of procedures to determine the level of automation required, given the inimitable human skillset. The current paper presents a human factors analysis to display the importance of tacit knowledge when introducing automation to optimise workforce sustainability and operators' psychophysical work needs and requirements. The Andreu World use case demonstrated two manufacturing processes: sanding and painting. Tacit knowledge and task duration was captured during both processes using eye tracking technology and semi-structured interviews. The results found that tacit knowledge was required for both processes, however, were utilised in separate ways by conducting a task analysis which specified performance-based levels. The painting process relied on visual inspection and procedural knowledge, whilst the sanding process relied on tacit and procedural knowledge. Hence, indicated that various levels of automation were required for each process to maximise human skillset and capability. The observations display the usage of tacit knowledge for everyday manufacturing tasks that provide a measure of how much automation is required within processes. Nevertheless, the paper paves a way for the factories of the future to focus on the implementation of automation with human skillset at the core of assembly lines.

### Introduction

The modern-day manufacturing environment has seen a significant increase in robotics. Specifically, industrial collaborative robots which are cyber-physical systems that allow safe interaction alongside a human operator in a shared workspace (Gualtieri et al., 2020). This partnership expands the potential application of robotic automation by combining machine strength with inimitable human skill. Further automation advancements display a great trajectory of growth for Industry 5.0; however, it is also essential to consider the extent to which collaborative systems are implemented with a human-centred focus to understand its impact on employee skills and needs. Although collaborative robotics (cobotics) aims to improve operator stresses and workforce sustainability (Correia Simões et al., 2020; Faccio et al., 2023), an influx can equally be argued detrimental to worker wellbeing, cognitive load, and skillset.

Collaborative robots frequently execute manufacturing-related tasks including assembly, automation, and high precision tasks (Jocelyn et al., 2017). These interconnected technologies

integrated in human-centred operations contribute to reduced physical strain, such as work-related biomechanical overload (Gualtieri et al., 2020), especially in manual assembly activities. This requires humans to deploy their explicit or tacit skill instead of focusing their efforts toward manual activities. Despite a reduced need for manual labour having the upmost advantage to physical health, we can also consider how the integrated cognitive load influences mental wellbeing. Currently, human task allocation is largely based on automation level and less on optimising human skillset. These differ between and within each industry (Moulières-Seban et al., 2017) as some processes may benefit from collaboration, however not all of them need to be collaborative.

Previous research demonstrates that the level of automation affects overall human-automation capability levels (Everitt et al., 2015). Without considering where operator skillsets are best placed to be most advantageous, companies risk a decrease in the productivity and efficiency of its workers. Furthermore, this shift presented to operator roles influences how one engages with their workload (Guerra-Zubiaga et al., 2020). Hence, it is important to consider how the operator role is optimised to not only manufacturing efficiency, but the wellbeing of its workers. This brings us to our research question; how do we decide which process fits for standard automation and where human robot collaboration is better?

To understand how best to deploy robotics so that we utilise human skills most effectively, it is important to integrate human factors. Human factors consider environmental, organisational, and individual characteristics that are vital to collaborative manufacturing assembly lines. One example of a factor includes mental workload which concerns the cognitive aspects related to work demand in response to collaboration at the workforce community level (Faccio et al., 2023). The sections below will define the main skills and knowledge involved in variety of manufacturing tasks in relation to a collaborative environment.

#### 1.1.1. *Tacit Knowledge*

Tacit knowledge is an umbrella term used to describe the intrinsic information and understanding experienced and learnt throughout any given role (Johnson et al., 2019). Tacit knowledge is acquired through contextual awareness which is harder to capture (Guerra-Zubiaga et al., 2020) and communicate as it is learnt through observation, imitation, and practice. Specifically, in manufacturing roles which require details of intricate processes and understandings. Tacit knowledge is valuable to companies as humans can further develop strategies (Johnson et al., 2019) by utilising it to combat task complexity, decomposition, and protocols. With the introduction of collaborative robots, organisations aim to rely less on the well-established experience and skills of long-term workforces and, instead, rely more on using the capabilities of available less-experienced personnel. A workforce with little tacit knowledge reduces the ability to independently overcome challenges, decision making and create evolved strategies (Johnson et al., 2019). This poses a potential threat to workforce effort, task difficulty and overall skill retention (Tadić Vujčić et al., 2017) which are essential to ensure that the needs of workers are retained such as autonomous work motivation.

#### 1.1.2. *Visual Inspection*

Visual inspection (VI) relies on tacit knowledge and refers to the human inspection of product quality and conformance to ensure outputs are completed to a high standard (Johnson et al., 2019). This traditional manual activity is central to acknowledging fault, injury, or failure (See et al., 2017). Tacit knowledge provides information on the standards required, as well as acknowledging how faults occurred in the first place. Humans are flexibly able to adapt to



changing environments whilst this poses as a challenge for robotics in ever changing industries. For example, sanding and painting tasks rely heavily on visual inspection to ensure that conventional robotics can achieve the same standard as humans. Although previous research has demonstrated that VI can be automated with IoT (Internet of Things) sensors (Gisginis., 2021), a study exploring VI capture found that working practices developed from tacit knowledge were necessary to overcoming task complexity (Johnson et al., 2019). This suggests that by displaying a more intuitive and in-depth "know-how" amongst manufacturing assemblies, operators can further advance mechanisms and processes which is more beneficial.

### 1.1.3. *Procedural Knowledge*

Procedural knowledge is included but not limited to falling under the tacit knowledge umbrella term. In comparison to tacit knowledge, which includes intuitions and hunches, it involves understanding how a task is completed with explicit knowledge through repetitive execution of activities (Pawlowsky, 2019). Procedural knowledge aids specific, finetuning details of tasks that cannot be integrated into AI (Artificial Intelligence) yet as it depends on a plethora of factors and circumstances (de Giorgio et al., 2020). A recent study has shown that system knowledge support is required among shop floor workers within production processes (Hoerner et al., 2023) as many operators continued to manually address anomalies or technique. This further supports that procedural knowledge promotes a more definite, clear response to challenges individuals may be met with (Hoerner et al., 2023), including sequential actions that follow production line processes (Johnson et al., 2019). Contrastingly, procedural knowledge also addresses sequential and repetitive tasks, whereby automation has proven most successful (de Giorgio et al., 2020). This allows for humans who can then focus on tasks that provide a higher added value for the company to experience more cognitively challenging tasks.

## 1.2. *The Present Paper*

The current study addresses a furniture manufacturer, Andreu World, within the AI PRISM project known as a use case. The use case consists of two processes – painting and sanding-where the introduction of AI assisted automation is considered. The current paper focus on the human factors analysis to provide the benchmark information for the technology development. Furthermore, it is the first step of end-user engagement to allow future co-creation activities between the different stakeholders involved in this use case. To achieve these aims the initial work conducted in this study will perform observations, interviews, and quantitative data collection to capture tacit knowledge and to propose how automation could be introduced to increase physical and psychological safety, increase productivity, and improve work ergonomics.

## **Methodology**

### 2.1. *Participants*

The study involved three participants: two operators working on the sanding process and one on the painting process. Participants were aware that they were being observed, hence an overt observation took place. All three participants had minimum of 5 years' experience on these processes.

## **2.2. Ethics**

This research was approved by the Cranfield University Research Ethics Committee, and conducted in accordance with the Cranfield Research Integrity Policy, the British Psychological Society's Code of Human Research Ethics, and the General Data Protection Regulation 2018.

## **2.3. Materials**

Semi-structured interviews: five questions were the focus of the interviews and then, depending on the answer, follow-up questions were asked. The structured questions were: 1. What is the most difficult during the painting/ sanding process; 2. Where potential errors can occur and how operators fix them; 3. What variation of the process does occur (novice vs experienced operators); 4. How long does the training take to be confident in the process? 5. Which aspects of the tasks are the most enjoyable?

Eye tracking glasses: Participants' gaze was tracked via SMI Eye Tracking Glasses (SensoMotoric Instruments ETG 1.7). The eye-tracking data was analysed using SensoMotoric's BeGaze© eye-tracking analysis software, utilising the Area of Interest (AOI) semantic gaze mapping. However, after further investigation of the collected data it was decided to use the recordings for the behavioural coding of actions to allow in-depth understanding of the main motions used by operators, their frequency and duration. The analysis was performed with behavioural observation coding software BORIS (Friard & Gamba, 2016).

## **2.4. Procedure**

Upon being briefed on the data collection aims, participants were guided through the informed consent and signed that they voluntarily agree to take part in the study. Participants were then assisted in putting on the eye tracking glasses and began performing their tasks. Operators performed five cycles of the painting process and for the sanding process operators finished one product (chair) each. The researcher was observing and noting down the actions completed by the operators.

Two processes were chosen by the Andreu World for the AI-PRISM project to consider: the chair painting and the chair sanding. The painting process consists of an operator, using spray paint, covering all the required surfaces of the product (in this case the chair), applying the necessary layers until the same tone as the sample is achieved. There are several product/chair variants differing in shape or wood type. These characteristics must be considered while completing this process. The sanding process is completed between each stage of painting to remove any inconsistencies and make the product as uniform as possible. The motion range and tactile pressure depends on product characteristics (shape and wood type).

The final step of the procedure was semi-structured interviews where the five main questions were asked and followed up by some clarifications of the process depending on the researcher observations. Once the semi-structured interviews were completed, participants were debriefed, reminded about the data withdrawal procedures, and thanked for their participation.

## **Results**

Following sections will review the observational findings from investigated processes. The conclusions will be drawn from triangulated interview and behavioural coding data. The following study found:

- Painting and sanding require different amounts of tacit knowledge.

- A task analysis (Table 1) demonstrated that sanding involved more tacit knowledge as part of the process in comparison to painting.

**Table 1:** Hierarchical task analysis with performance level for painting and sanding processes (green highlighted sections indicate reliance on tacit knowledge)

Painting					Sanding						
HTA		Purpose	Cues	Decision	Performance Level	HTA		Purpose	Cues	Decision	Performance Level
0	Starting the computer					0	Equip tools				
1	Preparing the environment (paint etc.)	To ensure there is appropriate tools (gloves) and space	Visual	Is the environment safe to begin?	Rule, Knowledge	0.1	Apply protection to hands	To protect hands from damage	Tactile	Are hands protected?	Rule
2	Getting the spray	To equip oneself	Tactile	Are the appropriate tools available and ready to use?	Rule	1	Take sanding paper	To equip oneself	Visual		Rule
2.1	Removing the previous colour from spray with water	To ensure a clean and clear base	Visual	Is there any previous colour?	Rule, Knowledge	1.1	Measure the right amount	To ensure there is sufficient sanding paper to cover the surface area	Visual	Is there enough sanding paper?	Knowledge
2.2	Colour sample match	To ensure the colour is appropriate	Visual	Is the colour correct?	Rule, Knowledge	2	Sanding the chair on outside	To start on the outside of the chair	Tactile	Is the outside of the chair sanded?	Knowledge
2.2.1	Decision: Testing the colour match with sample	To compare each colour with the desired colour	Visual	Does the colour and sample match	Rule, Knowledge	2.3	Rotate upwards and/or Move around the chair physically	To reach all areas of the chair	Visual, Tactile	Does the chair require rotating or moving?	Knowledge
2.3	Top cover	To complete the top of the chair	Visual	Is the top covered?	Knowledge	2.4	Sand across in direction of furniture shape	To cover the chair methodologically	Visual, Tactile	Is all the outside of the chair covered?	Knowledge
2.3.1	Turning the chair upside down	To adhere to the entire chair	Visual, Tactile	Is the chair held correctly?	Knowledge	2.4.1	Apply appropriate pressure and stroke length	To sand the wood accurately and effectively	Tactile	Is any further pressure required?	Skill
2.3.2	Spray from top to bottom following the component shape	To ensure the entire chair is covered and can follow where has been covered	Visual	Has the spray covered all areas of the chair?	Knowledge	3	Turn chair upside down	To reach all areas of the chair	Visual	Can the sanding process be approached with a different angle?	Rule, Knowledge
2.3.3	Rotate the chair	To cover all sides	Visual, Tactile	Has the chair been covered on the other side?	Rule	3.1	Sand chair interior and repeat process	To begin the inside of the chair	Tactile	Is the inside of the chair sanded?	Skill, Rule, Knowledge
2.3.4	Repeat the steps until even colour matching sample is achieved (3 spins)	To ensure even colour matching	Visual	Is colour matching achieved? Depending on the wood type and shape of the product	Skill, Knowledge	3.2	Replace sandpaper if required	To ensure the sandpaper is not used	Visual, Tactile	Does the sanding paper need to be replaced?	Rule
2.3.5	Compare the legs with the colour sample	To check the colour sample match	Visual	Are the legs colour sample achieved?	Rule	3.2.1	Decision: sanding paper replacement	To ensure maximum efficiency of the sandpaper	Visual, Tactile	How used is the sanding paper?	Knowledge
2.4	Spray the chair legs	To complete each individual chair leg	Visual	Are the chair legs covered	Rule	4	Check for any inconsistencies in product smoothness	To check for any corrections according to wood type	Visual, Tactile	Are there any inconsistencies?	Skill, Rule
2.4.1	Spray top and bottom sections of legs, then middle	To ensure the top and bottom are completed first	Visual	Are the edges of the chair completed first?	Rule, Knowledge	4.1	If product inconsistencies are present, then sand over	To sand specific areas	Visual, Tactile	Can further sanding help fix any errors?	Skill, Rule
2.4.2	Following component shape move the direction of the spray when spraying	To complete the middle of the chair leg	Visual	Is the middle of the chair leg sprayed correctly?	Rule, Knowledge	4.2	Use paint if required	To use solution on protected surfaces	Visual	Is paint required?	Skill, Rule
2.4.3	Check even colour matching	To ensure that legs are completed evenly	Visual	Are the chair legs evenly covered?	Rule	4.2.1	Check brush is dry and equip paint	To ensure no previous liquid could damage the solution	Visual, Tactile	Is the brush appropriate for usage?	Rule
3	Repeating the processes and starting on next chair	To continue the process line	Visual, Tactile	What other chairs to spray?		4.2.2	Paint over area	To cover/ correct any errors	Visual, Tactile	Is paint/ solution required?	Rule
						4.3	Wipe over chair with cloth	To remove any sanding residue	Visual, Tactile	Is there any residue that can be removed?	Rule

### 3.1. Painting

In the painting stage, the operator is located in a designated painting section which is isolated from the rest of the factory (to contain the spray paint). This area is configured to receive

continuously and periodically the two lines that lead the products through the workshop, which are placed from the beginning of the Painting Line. The operator uses industrial paint spraying apparatus to spray an initial layer of paint on the product. The task is dependent on visual information and experience about the shape and form of the product to ensure the product is sufficiently covered and is sprayed with efficient movements. The observation of this task showed that due to the assembly line being low, the operator needs to regularly bend to reach some parts of the product. In the subsequent interview, the operator confirmed that this puts a strain on his back, but it does not cause long term discomfort. The main aspects new operators need to learn and be aware about are dependent on the experience: (i) the shape of the component and drip; (ii) different wood soaks up the paint differently, and therefore the spraying needs to be either lighter or repeated for several times for greater coverage. By performing a count, table 1 displays that there are 1 tacit knowledge (skill) requiring steps for the painting process.

### **3.2. Sanding**

In the sanding stage, there were multiple operators working at the same area. During the time of observation there were four operators, and two of them were interviewed and their eye-tracking data was collected during the sanding process. The collected data revealed that the process heavily depends on the visual inspection and tactile information to determine if the quality of the finished product is acceptable. However, the process to achieve this relies on array of movements. As the video data from eye tracking glasses has revealed, ten motions/steps performed for the observed product (Figure 1), the further analysis was performed by coding different hand motions used in the sanding process, including the count and time they were used (Figure 1). Although both operators were observed while completing sanding on different components (the shape differed), the main similarity was both processes used long strokes while sanding the most indicating the need to perform overall surface sanding with light pressure. By performing a count, Table 1 displays that there are five tacit knowledge requiring steps to the sanding process.

Furthermore, plotting the actions against the time indicated how the different steps and motions are distributed over time and what is needed at the beginning of the process and how it evolves nearing the completion of the component. This data is plotted in Figure 1. Operators start with long strokes (blue colour on the figure) to cover all the surface area of the chair they are working on, with short strokes (magenta colour) left for the corners and where the parts are attached. However, over time and nearing the product sanding completion operators use more tactile sensing (1 finger (light pink colour on the figure), two fingers (grey colour) or with the palm (moss colour) – depending on the surface area where the imperfections are potentially observed). The final stages of the completion consist of fixes of tool usages (sanding stones for greater imperfections, scraping of bumps, and touching up the fixed areas with colour to enable even painting during the next stage of the assembly process).

Even though the process heavily relies on tactile information and visual inspection, the operators discussed that experience and knowledge about the wood type, different cuts, and even colours is essential. In fact, both operators previously have worked on antique furniture restoration, and they indicated that knowledge about the wood is important in the current job. Discussion on the training of new employees, operators indicated that if there is only one component, the training does take two to three months to learn, however, as the company has hundreds of products, the learning process can take a year or more. Operators indicated that the main priority in this station is the quality of the product. Some components can take 30 minutes, others 60 minutes, however, the most difficult aspect is knowing when to stop and ensuring that

high quality is achieved consistently. Considering the difficult aspects of the process both operators indicated that corners and joints are the most difficult as there is harder to achieve even surface sanding, and depending on the polish they might need use additional tools (scrapping) to remove the excess material. In the figures these areas are indicated by greater short stroke motion usage, where operators were applying greater force or smaller surface area.

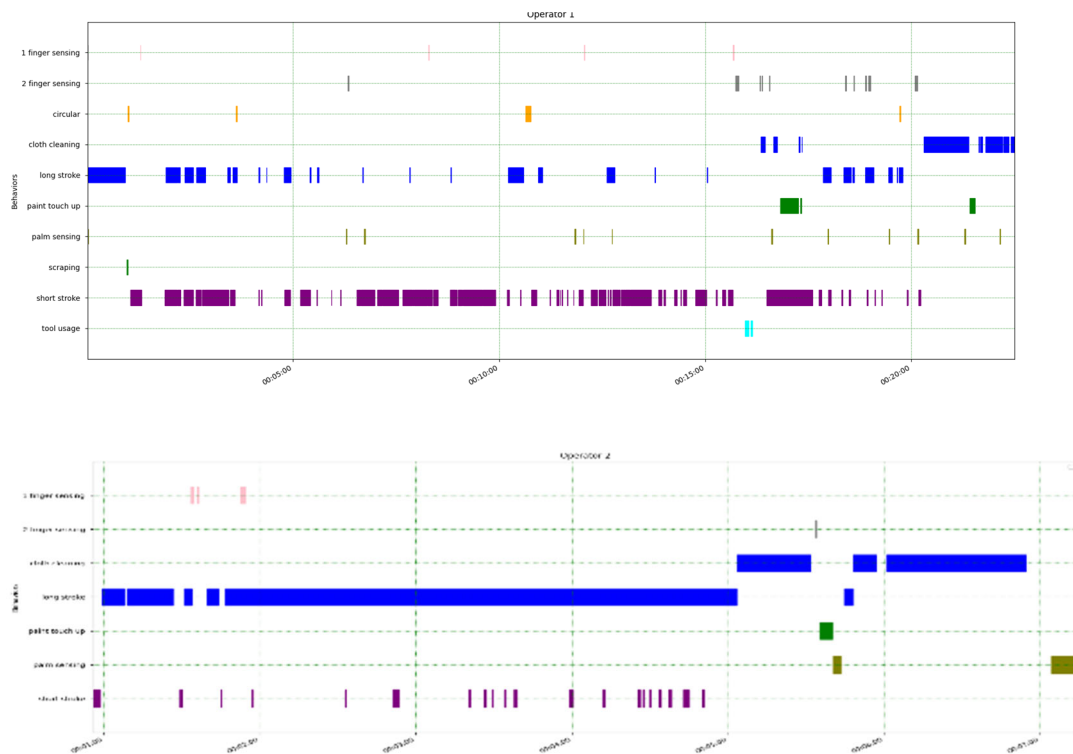
## Discussion

The use case demonstrates that tacit knowledge is a vital component to semi-automated processes. The behavioural coding of video data for eye-tracking glasses and short semi-structured interviews demonstrated that both the painting and sanding processes used tacit knowledge differently.

With the new founding popularity of collaborative robots in Industry 5.0, there are many benefits to consider. For example, emphasis on less productive motor actions enabling workers to focus on work planning activities, increased productivity, and less cognitive load (Zhu et al., 2020). However, to achieve these benefits, the research demonstrates considerations for planning new processes is needed as not all processes can or should be collaborative. Integrating automation can be challenging as both processes rely differently on automation and skill. The sanding process relied heavily on different motions for different periods of time (Figure 1) depending on tacit knowledge such as wood type, paper type and pressure required. With this type of knowledge integrated into a collaborative environment, workplaces are required to preserve and upkeep the understandings behind specific skills. Previous research supports this by illustrating the risks of further technological advancements causes the de-emphasising of tacit knowledge (Johannessen et al., 2001). This allows companies to direct their focus towards how tacit knowledge can be transferred effectively to others to expand the integration. Figure 1 shows initial sanding steps are the most time consuming and physical for the operator, and yet, they do not require lots of skill. Therefore, they could be performed with AI assisted automation. However, further sanding steps require almost all the skills described in the introduction, and thus cognitive flexibility, and should be left for the human to complete.

Furthermore, it should also be considered that automation aims to improve workers physical and emotional wellbeing to encourage workforce sustainability. This use case displays that the analysis of potential costs and gained benefits suggest that the painting process is most suitable for automation with human oversight due to physical comfort and procedural knowledge used in the task. However, sanding process incorporates a combination of more dexterous movement and engagement of tacit knowledge. Combination of human performing initial steps allowing the operator focus on critical aspects requiring their experience would increase physical and psychological wellbeing. Therefore, for this process human-in-the-loop process would be the optimal solution.

Future research may consider introducing psychometrics to further understand worker cognitive load such as the Mental Workload NASA (National Aeronautics and Space Administration) Task Load Index (Hart & Staveland, 1988). The scale may require revisions to suit the context of work; however, it provides a description into cognitive and manual control tasks. For example, the temporal demand of an integrated task may vary within the sanding process which requires tacit knowledge. This provides a benchmark as to how workers perform and feel about their tasks and provide key indicators of inconsistent functioning where inimitable skill is required. Focus groups also provide additional context to the tacit knowledge used to further understand operator perception. This is crucial to further understanding the transparency and transfer of tacit knowledge.



**Figure 1:** Time Spent on Motion Behaviours Project During Sanding Process

#### 4.1. Conclusion

The current research sought to shed light on the importance of tacit knowledge when integrating automation from a human factor's perspective. The use case demonstrated that certain processes thrive when relied on automation whilst others require inimitable human input. The impact of this research reveals the importance of how workplaces integrate their skills with automation and processes must be in place to achieve an overall sustainable workforce. Future work intends to have a greater sample size to complete a benchmark analysis and user-centered workshops to increase acceptance and engagement involving new processes and technology that reflect user needs and requirements.

#### ACKNOWLEDGMENT

This research has been supported by the EU project “AI-PRISM: AI-Powered Human-Centred Robot Interactions for Smart Manufacturing”. This project has received funding from the European Union’s Horizon Europe research and innovation program under grant agreement No 101058589. The authors would like to thank all consortium members for their contribution.

#### References

Correia Simões, A., Lucas Soares, A., & Barros, A. C. (2020). Factors influencing the intention of managers to adopt collaborative robots (cobots) in manufacturing

- organizations. *Journal of Engineering and Technology Management*, 57, 101574.  
<https://doi.org/10.1016/j.jengtecman.2020.101574>
- de Giorgio, A., Lundgren, M., & Wang, L. (2020). Procedural knowledge and function blocks for smart process planning. *Procedia Manufacturing*, 48, 1079–1087.  
<https://doi.org/10.1016/j.promfg.2020.05.148>
- Everitt, J., Fletcher, S., & Caird-Daley, A. (2015). Task analysis of discrete and continuous skills: A dual methodology approach to human skills capture for automation. *Theoretical Issues in Ergonomics Science*, 16(5), 513–532.  
<https://doi.org/10.1080/1463922X.2015.1028508>
- Faccio, M., Granata, I., Menini, A., Milanese, M., Rossato, C., Bottin, M., Minto, R., Pluchino, P., Gamberini, L., Boschetti, G., & Rosati, G. (2023). Human factors in cobot era: A review of modern production systems features. *Journal of Intelligent Manufacturing*, 34(1), 85–106. <https://doi.org/10.1007/s10845-022-01953-w>
- Friard, O., & Gamba, M. (2016). BORIS: A free, versatile open-source event-logging software for video/audio coding and live observations. *Methods in Ecology and Evolution*, 7(11), 1325–1330. <https://doi.org/10.1111/2041-210X.12584>
- Gisginis, A. (2021). *Production line optimization featuring cobots and visual inspection system*.
- Gualtieri, L., Palomba, I., Merati, F. A., Rauch, E., & Vidoni, R. (2020). Design of Human-Centered Collaborative Assembly Workstations for the Improvement of Operators' Physical Ergonomics and Production Efficiency: A Case Study. *Sustainability*, 12(9), 3606. <https://doi.org/10.3390/su12093606>
- Guerra-Zubiaga, D. A., Nasajpour-Esfahani, N., Siddiqui, B., & Kamperman, K. (2020). Implementing a Novel Framework to Create Tacit Knowledge Models to Support Human-Robot Interactions. *Volume 2B: Advanced Manufacturing*, V02BT02A013.  
<https://doi.org/10.1115/IMECE2020-23803>
- Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In *Advances in Psychology* (Vol. 52, pp. 139–183). Elsevier. [https://doi.org/10.1016/S0166-4115\(08\)62386-9](https://doi.org/10.1016/S0166-4115(08)62386-9)
- Hoerner, L., Schamberger, M., & Bodendorf, F. (2023). Using Tacit Expert Knowledge to Support Shop-floor Operators Through a Knowledge-based Assistance System. *Computer Supported Cooperative Work (CSCW)*, 32(1), 55–91. <https://doi.org/10.1007/s10606-022-09445-4>
- Jocelyn, S., Burlet-Vienney, D., & Giraud, L. (2017). Experience Feedback on Implementing and Using Human-Robot Collaboration in the Workplace. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 61(1), 1690–1694.  
<https://doi.org/10.1177/1541931213601911>
- Johannessen, J.-A., Olaisen, J., & Olsen, B. (2001). Mismanagement of tacit knowledge: The importance of tacit knowledge, the danger of information technology, and what to do about it. *International Journal of Information Management*, 21(1), 3–20.  
[https://doi.org/10.1016/S0268-4012\(00\)00047-5](https://doi.org/10.1016/S0268-4012(00)00047-5)
- Johnson, T. L., Fletcher, S. R., Baker, W., & Charles, R. L. (2019). How and why we need to capture tacit knowledge in manufacturing: Case studies of visual inspection. *Applied Ergonomics*, 74, 1–9. <https://doi.org/10.1016/j.apergo.2018.07.016>
- Moulières-Seban, T., Bitonneau, D., Salotti, J.-M., Thibault, J.-F., & Claverie, B. (2017). Human Factors Issues for the Design of a Cobot System. In P. Savage-Knepshild & J. Chen (Eds.), *Advances in Human Factors in Robots and Unmanned Systems* (Vol. 499, pp. 375–385). Springer International Publishing. [https://doi.org/10.1007/978-3-319-41959-6\\_31](https://doi.org/10.1007/978-3-319-41959-6_31)
- Pawlowsky, P. (2019). *Wissensmanagement*. De Gruyter.  
<https://doi.org/10.1515/9783110474930>

- See, J. E., Drury, C. G., Speed, A., Williams, A., & Khalandi, N. (2017). The Role of Visual Inspection in the 21<sup>st</sup> Century. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 61(1), 262–266. <https://doi.org/10.1177/1541931213601548>
- Tadić Vujčić, M., Oerlemans, W. G. M., & Bakker, A. B. (2017). How challenging was your work today? The role of autonomous work motivation. *European Journal of Work and Organizational Psychology*, 26(1), 81–93. <https://doi.org/10.1080/1359432X.2016.1208653>
- Zhu, Q., Wei, P., Shi, Y., & Du, J. (2020). Cognitive Benefits of Human-Robot Collaboration in Complex Industrial Operations: A Virtual Reality Experiment. *Construction Research Congress 2020*, 129–138. <https://doi.org/10.1061/9780784482858.015>



## **DEVELOPING UNMANNED AERIAL ROBOTICS TO SUPPORT WILD BERRY HARVESTING IN FINLAND: HUMAN FACTORS, STANDARDS AND ETHICS**

SARAH FLETCHER, ANNE-MARIE OOSTVEEN

*Cranfield University, Cranfield, Bedfordshire, UK*  
E-mail: [s.fletcher@cranfield.ac.uk](mailto:s.fletcher@cranfield.ac.uk) [J.M.Oostveen@cranfield.ac.uk](mailto:J.M.Oostveen@cranfield.ac.uk)

PAUL CHIPPENDALE

*FBK - Bruno Kessler Foundation, via Sommarive 18, 38123 Povo-Trento, Italy*  
E-mail: [chippendale@fbk.eu](mailto:chippendale@fbk.eu)

MICAEL COUCEIRO

*Ingeniarius Ltd, Rua Nossa Sra. Conceição 146, 4445-147 Alfena*  
E-mail: [micael@ingeniarius.pt](mailto:micael@ingeniarius.pt)

LAURA SMITH BALLESTER

*Universitat Politècnica de València, Valencia, Spain*  
Email: [lausmiba@ai2.upv.es](mailto:lausmiba@ai2.upv.es)

Wild berry picking plays a vital role in the Nordic diet, local culture, and regional economies. Despite the abundance of natural berry crops and their recognized health benefits, only a small fraction of the potential annual harvest is collected due to labour-intensive and imprecise harvesting methods, and the younger generation are not taking up this traditional activity. This means there is potential for commercial expansion. However, berry picking is arduous work involving significant physical and cognitive demands that may impact on both performance and wellbeing, and it can be hazardous work because of the challenges due to terrain and locating capabilities. To address these challenges, the FEROX project aims to revitalize the practice by leveraging AI, data, and robotics to improve working conditions and attract a broader demographic. UAV ‘drones’ are being equipped with advanced sensors to generate comprehensive visual data of berry-rich locations, and to develop 3D models of forests that can be viewed on mobile apps to provide accurate information on berry locations, abundance, and ripeness, which will enhance pickers’ performance. In addition to enhancing the picking experience, these systems will enable periodic monitoring of pickers to ensure wellbeing and provide assistance when needed, thereby enhancing safety. These advancements are expected to promote sustainable berry harvesting and unlock new commercial opportunities. However, by integrating key human factors and psychological analysis into system design that will bring improvements in working methods and conditions the project is aiming for significant positive impacts to berry pickers’ wellbeing and satisfaction. Moreover, this work is adopting a continuous ‘ethics by design’ approach to ensure ethical standards throughout the system lifecycle using the new robot ethics risk assessment protocol in current standards.

## 1. Background

### 1.1. *Wild food harvesting in Europe*

Foraging for wild food has an endemic and enduring history in Europe. It is one of the oldest forest activities in many countries, particularly in Scandinavia where fertile growing conditions are provided by long hours of sunlight in summer and large expanses of undisturbed wilderness. Numerous edible wild plants, berries and mushrooms are widely available for anyone to harvest thanks to common access and harvesting rights. Berries also require no real cultivation efforts due to their natural perennial wild growth. Berry abundance and availability means these inexpensive wild foods are a mainstay of traditional Nordic cuisine. Berries are primarily harvested by the general public for household consumption and for commercial sale (Turtiainen and Nuutinen, 2012; Kilpeläinen et al. 2016).

#### 1.1.1. *Commercial harvesting*

Commercial wild food harvesting is undertaken on a much smaller scale than domestic picking, but there are clear indications that there is room for international markets to be expanded. For some time, a ‘Nordic Diet’, which focuses on consumption of berries (along with vegetables, pulses, whole grain cereals and fish), has been associated with many health benefits, e.g. reducing diabetes, cardiovascular disease, cancer and obesity. Based on the consistency of these research findings, the World Health Organization officially advocated this type of diet in 2018, and this was widely publicised in mainstream media. As a result, local demand for berries and related food supplements increased in Finland (Ristioja, 2018). Similarly, greater public interest and demand has also emerged in recent years as a result of some berries being identified as ‘superfoods’ with similar health benefits (Markgren and Walldén Cerna, 2022), encouraging them to be added to new gastronomic menus around the world (Łuczaj & Pieroni, 2016). International consumer demand will likely boost levels of harvesting and exports, especially as climate change continue to threaten the yields of some traditional crops (Roitsch et al., 2022).

To meet the increasing demand for wild berries, berry trading companies are employing large numbers of foreign pickers for commercial picking. Since 2005, a large cohort of pickers from Thailand have been permitted to harvest berries in Finland, at a time which coincides with monsoons in their country. As a result of the migrant workforce, more produce is being placed on the market (Ristioja, 2018) and now only around a quarter of harvested berries are now being picked by Finns (Turtiainen and Nuutinen, 2021). There is considerable room to exploit and expand the commercial harvesting of berries given that there is increasing international demand but most are being consumed locally (Ristioja, 2018), and approximately 90–95% of crops are typically left unpicked each year (Paasilta et al., 2009).

#### 1.1.2. *Community harvesting*

By far, most wild food harvesting in Nordic countries is undertaken by members of the general public as recreational activity and for personal consumption (Mattalia et al., 2023). Forest foraging is an historic and sociocultural tradition, enabled by public access rights that allow anyone to walk and pick crops almost anywhere in the countryside, not only for personal use, but also for selling on. Small-scale domestic markets utilise locally harvested crops of around 4 million kg annually (Saastamoinen and Vaara, 2015). This local trade not only supports citizens’ personal incomes but also rural economies in general (Turtiainen and Nuutinen, 2012).

In a study of wild-food harvesting in 17 European nations<sup>1</sup> it was found that half of the rural populations across these countries were actively harvesting their own wild food<sup>2</sup> (Schulp et al., 2014). Based on these results, the authors of this study estimated that around 14% (65 million) of all EU citizens are likely to occasionally forage. In another study of foraging activity in Finland, it was found that around 56% of citizens go out to pick wild berries at least seven times each summer, irrespective of their socioeconomic status, and 87% of women aged 60-74 go out to forage (Korpela, 2013). In a good year, the total estimated amount of Finnish wild berry yield could be more than 500 million kilograms, with most (>70%) harvested for household use.



### 1.1.3. *Challenges of berry-picking in Finland*

In 2020, local and foreign pickers in Finland combined earned €19 million from harvesting wild berries (Euronews, 2021). The income is particularly rewarding for migrant workforces. In 2021, the two-month salary of a prolific picker equated to 15 years of income in Thailand (YLE, 2021). However, such a high income in such a short amount of time comes at a price as working conditions are very challenging. Berry pickers need to work as efficiently as possible given that their earnings depend on the type, quality, and quantity (weight) of berries that they harvest. This means their working day is long, physically strenuous, and mentally demanding.

Firstly, the working day for berry pickers is long so that they can make the most of their time in the forest, typically rising early (e.g. around 4.30 am) to travel some distance by road to the chosen forest location, where they will then walk and forage throughout the day, often with little or no breaks and not returning back to base before 8.00/9.00 pm.

Secondly, the physical strenuousness of fruit picking work is high, as pickers need to constantly move, bend and stretch over rough terrain in order to reach and gather berries using combing / cutting tools which are then emptied into large buckets that need to be carried. Lifting and carrying these buckets obviously gets more taxing through the day as weight increases, and

<sup>1</sup> Austria, Belgium, Bulgaria, Cyprus, Czech Republic, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, Ireland, Italy, Latvia, Lithuania, Luxembourg, Malta, Netherlands, Poland, Portugal, Romania, Slovakia, Slovenia, Spain, Sweden, United Kingdom.

<sup>2</sup> Berries, plants, mushrooms and game (wild animals and birds).

as pickers inevitably become increasingly fatigued. All this exertion is exacerbated by the warm, highly humid conditions of the rainy forests during the summer season.

Thirdly, cognitive demands are a less salient but nonetheless significant challenge in berry picking work. Pickers need to accurately identify the most suitable areas to forage, and how to reach ripe berries. To identify the most lucrative locations to forage is a considerable challenge even for local pickers. Prime crop locations are estimated using personal experience of weather and geographical conditions. The migrant workforce must rely on the accuracy of this speculative knowledge being provided to them by locals or try to identify prime picking areas themselves. Larger numbers of foreign pickers in recent years means that the overall workforce has become less equipped with such local / experiential knowledge (Tikkanen, 2015). Acquiring knowledge may be especially problematic for foreign pickers due to language limitations. One of the navigational aids that they use is to add markings on trees to help them recognise traversed paths and find their way back. However, if a worker becomes lost or compromised due to injury these markings would not help. Pickers may also use their own smartphones to locate themselves in the forest using online map services. However, these maps provide insufficient detail, network coverage is often poor in the forests, and there is a risk that the device may be dropped, malfunction, or run out of batteries. Hence, pickers' cognitive demands are also heightened by this need to maintain situational awareness and avoid personal injury as they negotiate unfamiliar and uneven terrain. Several pickers get into difficulty each year, becoming lost or stuck (YLE, 2020).

## **1.2.     *Unmanned aerial vehicles***

Unmanned aerial vehicles (UAVs) or 'drones' are a frequently studied area of robotics, as they offer unique manoeuvrability, hovering, and low altitude flight capabilities for a diverse range of applications and tasks (Valavanis, 2017). Without the constraint of a human payload, UAVs can carry a wide variety of precision sensing equipment over wide areas of interest for long periods, effortlessly gathering Gbytes of observational data. The major drawback is that the majority of low-cost UAVs are designed for human operation and those with autonomous function must by law be restricted to Visual Line of Sight (VLOS) flights.

### **1.2.1.   *Applications for work environments***

UAVs are becoming an increasingly attractive technology for supporting manual work across many industrial sectors, especially suitable for search and inspection tasks. In recent years there have been significant advances in on-board sensing, batteries, navigational controls, and predictive flight models, which means that they can now operate with increasing levels of automation (Kingston et al., 2016). However, they are still a tool which must only be used directly alongside or in partnership with human operators.

The construction industry is a clear leader in adopting UAVs to join workforce tasks such as site surveying, mapping, monitoring, inspection and maintenance (Jeelani and Gheisari, 2022). UAVs are now also starting to be deployed in other sectors and applications such as manufacturing (Maghazei et al., 2020), mining (Shahmoradi, et al., 2020), rail (Golightly et al., 2020), agriculture (van der Merwe et al., 2020), healthcare (Hiebert et al., 2020), humanitarian operations (van Wynsberghe and Comes, 2020), and many more.

Agricultural UAV applications are primarily deployed in large scale commercial farming activities such as crop and livestock monitoring and chemicals distribution (van der Merwe et al., 2020). Up to now, there has been little or no research and development of UAV systems to support wild food foragers, either commercial or recreational. UAVs are ideal for the

photogrammetric mapping of crops when flights are unhindered by obstacles, but in forest mapping scenarios, under-canopy flights between trees and foliage either with a human pilot or AI navigation is currently very challenging.

In addition to the many technical challenges faced by drone deployment in berry picking environments, little research has been conducted to explore the collaborative human-centred aspects of drone-supported berry picking, and whether this proposition can deliver measurable improvements to pickers' productivity, health, and wellbeing.

#### 1.2.2. Human-UAV collaboration

The field of human-UAV collaboration focuses on “*evaluating and developing new control modalities, designing new applications where humans interact with drones, and enhancing such interaction by understanding how humans perceive the interaction*” (Tezza and Andujar, 2019, p.167439). Human roles are determined by the level of autonomy and application of the UAV system. They can operate the UAV using a control interface, act as a supervisor with the ability to take control if necessary, collaborate with the UAV as a task partner, or simply be a recipient without control authority (Tezza and Andujar, 2019). As UAV systems become more autonomous, operators take on higher-level supervisory roles, but successful task completion still relies on human-robot collaborations.

Human-robot interaction research has provided valuable insights that have guided the technical design of collaborative robotic systems. A small number of studies have also identified associations between specific robot characteristics and human responses that directly affect the effectiveness of collaborations. For example, we know that human trust (that the robot will make appropriate decisions and behave in an expected and understandable manner) is crucial for effective collaboration (Gil et al., 2019), and levels of trust are influenced by specific features such as perceived robot speed, reliability and safety (Charalambous et al., 2016). However, UAVs possess unique characteristics that require a distinct dimension of human interaction and collaboration research. Current knowledge about the impacts of UAV characteristics on the responses of human collaborators, operators and other individuals in the vicinity of UAVs is unknown. To address this need, the field of “human-drone interaction” (HDI) has been proposed as a dedicated area of investigation for understanding, designing, and evaluating drone systems for human users (Tezza and Andujar, 2019). Moreover, incorporating user-centred design principles and empirical ergonomic data into multi-modelling techniques can greatly enhance the underlying models guiding UAV functions (Golightly et al., 2020).

In addition to trust, human-robot collaboration is also influenced by a range of specific psychological and affective states that the human experiences in a situation or context, and which can be highly instrumental to performance and wellbeing outcomes. In particular, mental workload, situation awareness and (intrinsic) job satisfaction are key factors that are directly influenced by system characteristics such as task complexity, level of autonomy, speed and safety (perceived and actual), etc. The interplay of these human factors and UAV system characteristics has not yet been defined and remains a research gap.

A range of ethical issues are highly important to the design and deployment of UAVs for human collaboration, particularly concerning privacy and personal data security. To address these concerns, various guidance documents and ethical risk assessment protocols are emerging. Amongst these, the world's first formal standard for ethical design and application of robots including UAVs (BS 8611:2016) provides general guidelines on safe design, protective measures and how to conduct ethical risk assessments. This guidance emphasizes the importance of transparency, accountability, and the consideration of social, legal, and ethical

factors throughout the UAV lifecycle. Ethical risk assessment protocols, such as those derived from the Responsible Research and Innovation (RRI) approach, aim to identify and mitigate potential ethical risks associated with UAV use, ensuring responsible and socially acceptable deployment. By integrating ethical considerations into UAV operations, these literature, guidance, and assessment protocols contribute to the development of ethically sound practices and foster public trust in the use of UAV technology (Torresen, 2018).

### **1.3. *Summary of the problem***

Wild berry harvesting is clearly important to the local culture, economy and nutrition in Finland and other Nordic countries, and of increasing commercial importance in response to international demand. There is a clear potential for significantly expanding current harvesting levels, which would be advantageous for local pickers and migrant workforces alike, as berry picking work is arduous and potentially hazardous. UAV technologies now offer the capabilities needed for assisting human workers, particularly in navigation and monitoring, but their effectiveness will rely on user-centred and ethical design. Thus, UAV services must be developed in harmony with human-centred design principles.

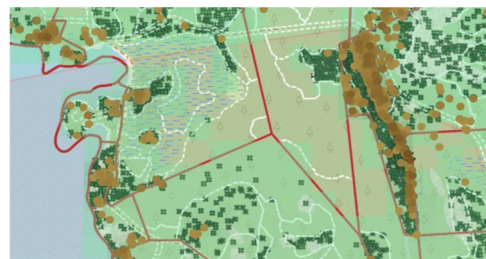
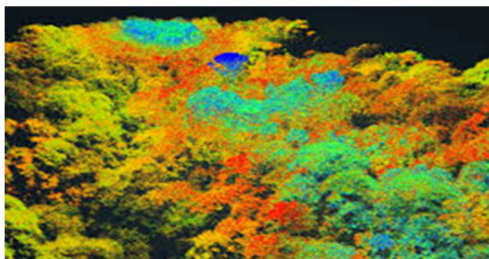
## **2. FEROX**

FEROX (Fostering and Enabling AI, Data and Robotics Technologies for Supporting Human Workers in Harvesting Wild Food) is a European Commission Horizon Europe research project, that aims to develop a UAV solution for wild berry picking in forest. FEROX will exploit human-centred and ethical principles in its design process to create an appropriate, collaborative and usable robotics, AI and data science system.

### **2.1. *Project aims and objectives***

A fully automated solution for monitoring and harvesting wild berries from forests is not going to be possible for many years, if ever, and is ultimately an ultimate target for commercial enterprises. Berry sensing, terrain navigation and fruit manipulating are currently essential human qualities that will be extremely difficult to replicate by a machine. Moreover, full automation may never be desirable if it conflicts too much with local conventions and culture.

Considering current technological capabilities and sociocultural limitations, FEROX is exploring the use of a variety of lightweight and heavy-lift UAVs to locate berry crops in forests and then help human pickers to carry their harvests to centralised vehicles. FEROX will improve crop location estimates by integrating country-level airborne laser scanning (ALS) data to produce detailed AI-driven crop yield maps that include key terrain features such as ditches, small ponds, and individual trees. These maps will be fused with frequent UAV observations to make positioning and navigation much easier and pickers will be able to familiarise themselves with surrounding terrain and find their way out of undesired areas like swamps.





In addition to these technical design goals, FEROX aims to enhance the performance and wellbeing of fruit pickers, by alleviating some of the major challenges and risks that they currently face. Berry pickers' safety and trust will not only be enhanced through the supply of more reliable and accurate navigational data, but also by a novel 'WatchDog' mode that will allow them to be periodically observed to verify their supposed location and wellbeing. The UAVs will be able to detect and review potential picker problems and to provide appropriate support in response. To enhance pickers' acceptance of the FEROX solutions, and thereby achieve performance and wellbeing improvements, the project includes a strong element of human analysis to identify context-specific user requirements. This information will be exploited to inform system and interface design and help evaluate the impact of the new technology on performance and wellbeing.



## 2.2. *Human analysis*

The methodology for human data collection and analysis will address three principle objectives: to identify user requirements, inform system design, and evaluate impacts.

### 2.2.1. *Identify user requirements*

Identifying what the intended users of a new technology want or need is a foundational step in effective system design. In the specific context of developing UAVs to optimise berry picker performance and wellbeing, where there are currently no directly relevant research findings to guide design, the first task for FEROX human analysis has to be to identify their requirements. To this end, the experiences and expectations of berry pickers will be investigated to establish conventional work practices and identify specific limitations and problems, and to explore how they believe a UAV system should be designed to best assist them and identify preferences and needs. To gather this information, volunteer samples from the berry picking workforce in Finland will be recruited to take part in a set of data collection activities:

- i. One-to-one interviews designed to elicit personal experiences and opinions.
- ii. Online psychometric surveys designed to anonymously capture opinions and current levels of key factors (workload, satisfaction, job characteristics), and non-identifying demographic data (experience, nationality, age, etc.).
- iii. Eye tracking video data to document current manual berry picking techniques.

### 2.2.2. *Inform system and interface design*

Formal guidance to enhance the design of UAVs for human performance and wellbeing is also scarce. Several standards to guide functional safety and performance in UAV design exist; for example, there are fifteen technical standards developed by ISO Technical Committee 20, Sub Committee 6. In terms of human aspects, there are numerous standards containing general ergonomic principles, some of which could perhaps be relevant to the context of UAV design, such as interface design (ISO 9241-161:2016) or general work system design (ISO 6385:2016),

etc. However, there are no standards specifically to guide user-centred design of UAV systems. One recent technical report provides a first international standards document on human-centred aspects of robotics (ISO/TR 9241-810:2020) but this provides a general level of guidance that is not specific to UAV characteristics. Consequently, without dedicated standards or specifications for this context, the next project objective will be to apply the user requirements data to inform the design of FEROX systems and interfaces. Specifically, the information provided by interviews on user experiences and opinions, and the eye tracking data that shows real picking work being performed will be distilled and mapped to provide a usable set of guidelines.

### 2.2.3. *Evaluate impacts*

The final objective for the human analysis work in FEROX will be to measure the impact of the new UAV solutions, to assess how well the user-centred design has been successful in improving the work and assisting pickers. The performance and wellbeing data gathered in the user requirements phase will be followed up by the same or comparable techniques so that changes can be assessed qualitatively or measured quantitatively. Performance impacts may be identifiable in changed task techniques and procedures, productivity outcomes, limitations and problems, etc. Impacts on wellbeing may be more measurable via statistical comparisons of the results from the online psychometric surveys. Additional measures are likely to be introduced to strengthen the findings. For example, although it is not possible to measure human trust in UAVs within the first data collection phase – because systems are not yet implemented – reliable psychometric measures of human trust can be administered at a later stage when pickers have experienced the new solutions, and this will expand impact evaluations.

## 2.3. *Ethical analysis*

FEROX is committed to an ‘ethics by design’ approach in which ethical standards will be monitored and addressed throughout. The most prominent ethical issue to consider in the project concern is public privacy and data security. It is vital that pickers trust that the UAV systems are not collecting unnecessary or covert data, and that any legitimate and agreed data is stored and managed with utmost security, in accordance with the European General Data Protection Regulation (GDPR). This aspect of trust is particularly important for any pickers who may be anxious about the UAV surveillance but felt pressured to consent in order to earn money from their yields but is crucial for all who encounter the UAVs in the forests. It must be remembered that the UAVs will not only operate to monitor the location and wellbeing of berry pickers, their flights in the outdoor environment to which everyone has access means they will also be experienced by other people in the vicinity. Clearly then, the UAVs may affect people who have consented to, and are aware of, the surveillance, but also bystanders who are not forewarned and may not be as agreeable, so the project needs to consider how best to manage both user and non-user expectations.

With such a long history and high level of local / public participation in berry picking the potential impacts on local culture must also be considered. Technology development has (sometimes infamously) neglected consideration of impacts on the wider sociocultural environment and disrupted communities of people. However, as part of FEROX’s commitment to ethical standards, the human data collection and analysis will not only endeavour to capture the requirements and expectations of pickers / users but also the opinions and needs / preferences of local communities and wider stakeholders. As the UAVs will operate in the forests, potential impacts on the natural environment must also be reviewed and addressed. FEROX will monitor



and assess potential effects on local wildlife and agriculture in liaison with local experts to ensure the systems are designed to minimise negative impacts.

The ‘ethics by design’ approach that will be maintained throughout FEROX means that these ethical issues – and many others that are anticipated or may yet emerge as the project progresses – will be continuously monitored and addressed. Moreover, the objective of this approach is to ensure ethical principles last throughout the entire life cycle of the UAVs, to the eventual end and disposal of the systems. To this end, as part of the human analysis, ethical risk assessments will be made as an ongoing activity by applying the BS8611 evaluation protocol. This will be overseen by a designated formal Ethics Board consisting of key members of the project team along with a selected external expert Ethics Advisor expert.

### **3. Conclusions and future directions**

The development of unmanned aerial robotics for supporting wild berry harvesting holds significant potential for addressing the challenges faced by traditional harvesting methods. This paper has highlighted the importance of wild berry picking which, despite its cultural significance and increasing international demand, its current harvesting methods are still labour-intensive, mentally and physically challenging, and lack precision in locating berry crops. The FEROX project, utilizing advancements in AI, data, and robotics, aims to enhance the working methods and conditions of wild berry pickers by using autonomous drones, or UAVs, equipped with various sensors and accessible through intelligent mobile apps. This collaborative approach between humans and UAVs offers the opportunity to improve berry location accuracy, increase overall yield and incomes, and provide monitoring and safety benefits. Furthermore, the incorporation of robotics and technology into berry picking endeavours has the potential to engage younger generations and widen the demographic of pickers, creating a sustainable future for this tradition. Importantly, however, this project is not only integrating human factors and psychological science throughout, it is also maintaining an ‘ethics by design’ approach to uphold ethical best practice throughout and beyond the project as a full life-cycle commitment. Future reporting and dissemination of findings will provide information to guide other technology development projects on how to successfully incorporate human and ethical issues.

### **Acknowledgments**

This work is partly funded by the EU FEROX project (<https://ferox.fbk.eu/>). FEROX has received funding from the European Union's Horizon Europe Framework Programme under Grant Agreement No. 101070440. Views and opinions expressed are however those of the authors only and the European Commission is not responsible for any use that may be made of the information it contains. Cranfield University’s participation in the project is supported by funding from the UKRI.

### **References**

1. Adams, J. A., Humphrey, C. M., Goodrich, M. A., Cooper, J. L., Morse, B. S., Engh, C., & Rasmussen, N. (2009). Cognitive task analysis for developing unmanned aerial vehicle wilderness search support. *Journal of Cognitive Engineering and Decision Making*, 3(1), 1-26.
2. Charalambous, G., Fletcher, S., & Webb, P. (2016). The development of a scale to evaluate trust in industrial human-robot collaboration. *International Journal of Social Robotics*, 8, 193-209.

3. DiSalvo, C., & Jenkins, T. (2017). Fruit are heavy: a prototype public IoT system to support urban foraging. In *Proceedings of the 2017 Conference on Designing Interactive Systems* (pp. 541-553).
4. Euronews, 2021. Did Finland do enough to protect its foreign berry pickers from Covid? Accessed 10/03/23 at: <https://www.euronews.com/2021/09/16/did-finland-do-enough-to-protect-its-foreign-berry-pickers-from-covid>
5. Gil, M., Albert, M., Fons, J., & Pelechano, V. (2019). Designing human-in-the-loop autonomous cyber-physical systems. *International journal of human-computer studies*, 130, 21-39.
6. Finnish Food Authority, 2021. Marsi 2020 Luonnonmarjojen ja -sienten kauppantulomäärät vuonna 2020. (Marsi 2020, commercial collecting of natural berries and mushrooms in 2020). Online, retrieved 15/10/21 at: <https://www.ruokavirasto.fi/globalassets/viljelijat/tuet-ja-rahoitus/marsi-2020-raportti.pdf>
7. Golightly, D., Gamble, C., Palacin, R., & Pierce, K. (2020). Applying ergonomics within the multi-modelling paradigm with an example from multiple UAV control. *Ergonomics*, 63(8), 1027-1043.
8. Hiebert, B., Nouvet, E., Jeyabalan, V., & Donelle, L. (2020). The application of drones in healthcare and health-related services in north america: A scoping review. *Drones*, 4(3), 30.
9. Jeelani, I., & Gheisari, M. (2022). Safety challenges of human-drone interactions on construction jobsites. *Automation and robotics in the architecture, engineering, and construction industry*, 143-164.
10. Kingston, D., Rasmussen, S., & Humphrey, L. (2016, September). Automated UAV tasks for search and surveillance. In *2016 IEEE Conference on Control Applications (CCA)* (pp. 1-8). IEEE.
11. Kilpeläinen, H., Miina, J., Store, R., Salo, K., & Kurttila, M. (2016). Evaluation of bilberry and cowberry yield models by comparing model predictions with field measurements from North Karelia, Finland. *Forest Ecology and Management*, 363, 120-129.
12. Korpela, Jukka K. (2008), 'Jokamiehen oikeuksiin selvyyttä', in: *Oikeus* 37, 93-97
13. Łuczaj, Ł., & Pieroni, A. (2016). Nutritional ethnobotany in Europe: From emergency foods to healthy folk cuisines and contemporary foraging trends. *Mediterranean wild edible plants: ethnobotany and food composition tables*, 33-56.
14. Maghazei, O., & Netland, T. (2020). Drones in manufacturing: Exploring opportunities for research and practice. *Journal of Manufacturing Technology Management*, 31(6), 1237-1259.
15. Markgren, R., & Walldén Cerna, F. (2022). A comparative study of the wild berry industry's innovation system in Sweden and Finland: Conditions affecting the prospect of an industry development.
16. Paassilta M., Moisio S., Jaakola L. and Häggman H. (2009) Voice of the Nordic wild berry industry. A survey among the companies. Oulu University press, Oulu.
17. Ristioja, A. (2018). Luonnontuoteala. Accessed 09/06/23 at: <https://julkaisut.valtioneuvosto.fi/handle/10024/160687>
18. Roitsch, T., Himanen, K., Chawade, A., Jaakola, L., Nehe, A., & Alexandersson, E. (2022). Functional phenomics for improved climate resilience in Nordic agriculture. *Journal of Experimental Botany*, 73(15), 5111-5127.
19. Saastamoinen, O., & Vaara, M. (2009). Small-scale forest related activities in the everyday life of the Finns: results of time-budget studies. *Small-scale Forestry*, 8(4), 425-445.
20. Shahmoradi, J., Talebi, E., Roghanchi, P., & Hassanalian, M. (2020). A comprehensive review of applications of drone technology in the mining industry. *Drones*, 4(3), 34.
21. Tezza, D., & Andujar, M. (2019). The state-of-the-art of human-drone interaction: A survey. *IEEE Access*, 7, 167438-167454.
22. Tikkanen, P. (2015). Physical functioning among community-dwelling older people (Doctoral dissertation, Itä-Suomen yliopisto).
23. Torresen, J. (2018). A review of future and ethical perspectives of robotics and AI. *Frontiers in Robotics and AI*, 4, 75.

23. Turtiainen, M., & Nuutinen, T. (2012). Evaluation of information on wild berry and mushroom markets in European countries. *Small-scale Forestry*, 11, 131-145.
24. This is Finland (2013). Accessed 09/06/23 at: <https://finland.fi/life-society/treasures-of-the-boreal-forests/>
25. Valavanis, K. P. (2017, July). Unmanned Aircraft Systems challenges in design for autonomy. In 2017 11th International Workshop on Robot Motion and Control (RoMoCo) (pp. 73-86). IEEE.
26. van Wynsberghe, A., & Comes, T. (2020). Drones in humanitarian contexts, robot ethics, and the human–robot interaction. *Ethics and Information Technology*, 22, 43-53.
27. van der Merwe, D., Burchfield, D. R., Witt, T. D., Price, K. P., & Sharda, A. (2020). Drones in agriculture. *Advances in Agronomy*, 162, 1-30.
28. YLE 2020. Lapland police report busy summer rescuing lost berry pickers. Accessed 13/06/23 at: <https://yle.fi/a/3-11458780>
29. YLE, 2021. Osa thaipaimijoista keräsi 15 vuoden tulot reilussa kahdessa kuukaudessa. News 6 Oct 2021



**SECTION-4**  
**AI REGULATIONS**



## **AI, GLAM, INNOVATION AND TRUST: TOWARDS VALUE-BASED REGULATION BY DESIGN**

KELLY BREEMEN and VICKY BREEMEN

*Utrecht University, The Netherlands*

The EU policy agenda addresses the development and use of AI in various sectors, including environment and health, finance, mobility, agriculture, education and culture<sup>1</sup>. Recurring keywords are ‘trust’ and innovation versus safety and fundamental rights. The aim is for AI to be ‘human-centric’<sup>2</sup>. Our contribution zooms in on culture, and more specifically, on ‘Cultural AI’, i.e. “the study, design and development of socio-technological AI systems that are implicitly or explicitly aware of the subtle and subjective complexity of human culture”<sup>3</sup>, an aspect which so far has not received much attention in the policy discussions. It does so in the context of digitally unlocking cultural heritage via GLAM (galleries, libraries, archives and museums) institutions, since the systems and structures these institutions use to handle, organise and make accessible the materials they harbor and safeguard arguably originate from dominant perspectives and world views. Consequently, due to the resonance of power imbalance, GLAM institutions might even be regarded as exclusionary in nature. As it is argued that such dominant systems should give way to “alternative distribution of control”<sup>4</sup>, how can ‘Cultural AI’ contribute to “reevaluat[ing] the workflows and procedures of digital archiving and curation”<sup>5</sup>.

Although a recent study for the European Parliament examines opportunities and challenges of AI in the context of cultural heritage and museums, this study highlights only a number of topics, such as the use of AI for restoring or completing works, author identification or the detection of hidden archeological sites. Cataloguing and information management are merely mentioned briefly<sup>6</sup>. It is precisely this angle and gap that our contribution addresses, namely of how to benefit from the application of AI in the sphere of unlocking content online while being aware of the risks this technology poses, such as bias and lack of context-specificity of this technology. How can the values at stake be safeguarded?

To that end, our contribution, which is positioned at the interplay between law, culture and technology, takes an interdisciplinary law & humanities approach to critically assess theoretical frameworks, practical tools and fundamental questions surrounding the meaningful and culturally sensitive use of AI for GLAM practice, which we aim to concretise as value-based regulation by design for unlocking cultural heritage online. This approach should enable us to balance innovation with fundamental values such as access, stewardship, self-determination, representation, participation, fairness and trust.

-----  
<sup>1</sup> See for instance Proposal for a regulation of the European Parliament and of the Council laying down harmonized rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative Acts {SEC(2021) 167 final} - {SWD(2021) 84 final} - {SWD(2021) 85 final} Brussels, 21.4.2021 COM(2021) 206 final 2021/0106 (COD), p. 1.

<sup>2</sup> Idem

<sup>3</sup> Definition taken from <https://www.cultural-ai.nl/>.

<sup>4</sup> K. Christen & J. Anderson, ‘Toward slow archives’, *Archival Science* 19 (2019), p. 99.

<sup>5</sup> Christen & Anderson 2019, p. 92.

<sup>6</sup> M. Pasikowska-Schnass & Y-S. Lim, Artificial intelligence in the context of cultural heritage and museums. Complex challenges and new opportunities, European Parliamentary Research Service, April 2023, p. 2-3, 9.

## **PROOF OF CAUSATION UNDER UNCERTAINTY**

ELBERT DE JONG

*Utrecht University, The Netherlands*

In 2022, four European directive proposals saw the light of day that provide for different evidentiary rules in a variety of liability issues. The rules can be found in the proposed Directive amending the Industrial Emissions and Waste Directive, the proposed Directive on Air Quality and Cleaner Air for Europe, the proposed Directive on Product Liability and the proposed Directive on AI Liability. The common thread is that they address situations where there is scientific uncertainty about causation. The rules aim to (partly) eliminate three common causes of proof problems for injured parties: scientific uncertainty about the causal link, complexity of (connected) causes of damage and the existence of inequality in information positions between the defendant and the claimant. However, important questions arise about the effectiveness of and consistency between the various schemes, which will be discussed.



## **THE GOVERNANCE OF GENERATIVE AI: OBSERVABILITY, MODIFIABILITY, ACCESSIBILITY**

FABIAN FERRARI

*Utrecht University, The Netherlands*

The increasing permeation of society by generative AI systems, such as ChatGPT, has given rise to a pressing issue that largely remains unresolved: the need for governance mechanisms that ensure effective democratic oversight over those systems. To establish this oversight, it is essential that generative AI systems can be opened up for regulatory scrutiny. But transparency is not an end in itself. This talk argues that there are three additional conditions to ensure effective democratic oversight: analytical observability, technical modifiability, and public accessibility. Taken together, these conditions point to clear pathways to thwart the dominance of corporate oversight mechanisms by firms such as OpenAI, Microsoft, and Google. As generative AI systems infiltrate economic, political, and cultural interactions, the stakes for turning corporate oversight procedures over those infrastructures into democratic governance institutions are high.

## **PRIVACY ASPECTS OF AI REGULATION**

LESLEY BROOS

*Utrecht University, The Netherlands*

Big data processing and Artificial Intelligence are inextricably linked, and given the broad definition of personal data, the processing of personal data is almost inevitable when one is processing big data. In the past, many authors have therefore warned for the personal data protection risks that AI therefore harbors. So, unsurprisingly, the proposed AI act contains a number of arrangements that link up with the General Data Protection Regulation. Which way has the balance between AI regulation and data protection tipped? Is the AI act good or bad news for the protection of personal data? These and other questions at the interface of AI and data privacy are addressed in this presentation.

## **SECTION-5**

**WORKSHOP: HOW DO YOU WANT TO BE GOVERNED?  
SHOULD WE PACE INNOVATION? A CHATGPT CASE  
STUDY**



## HOW CAN WE COMPARE LLMS IN THEIR CURRENT STATE WITH THE HUMAN CAPABILITY OF UNDERSTANDING?

LUIZA ŚWIERZAWSKA  
Utrecht University, Utrecht, The Netherlands  
E-mail: [l.j.swierzawska@uu.nl](mailto:l.j.swierzawska@uu.nl)

The recent developments and growing presence of increasingly advanced machines have sparked debates on AI systems becoming human-competitive. This paper attempts to answer the question of how can we compare LLMs in their current state with human capability of understanding. It highlights the importance of an interdisciplinary approach to AI governance, encouraging the collaboration between philosophers and engineers.

### 1. Introduction

The recent developments and growing presence of increasingly advanced machines have sparked debates on AI systems becoming human-competitive. During the Philosophy of Deep Learning conference held by New York University in March this year, scholars have reflected on cognitive capabilities of deep neural networks. Given the challenges that the society is now facing, one of the lecturers, Cameron Buckner, stressed the relevance of philosophical debates between nativism and empiricism in engineering and designing core knowledge systems, such as LLMs.<sup>1</sup> Following this reasoning, this paper attempts to answer the question of *how can we compare LLMs in their current state with human capability of understanding?* It highlights the importance of interdisciplinary approach to AI governance, encouraging the collaboration between philosophers and engineers.

### 2. The Origins of Knowledge: Philosophical Debates

Philosophical discussions have long focused on the question of what understanding is and how we acquire it. Before engaging in their first hands-on experiences with the objects or subjects they are learning about, what knowledge do children already have?<sup>2</sup> Which aspects of knowledge remain constant throughout the course of human development, and which do not, beginning with early childhood when a person begins to understand their environment?<sup>3</sup> These questions have sowed the seeds of dialogue known as the intellectual debate between nativists and empiricists. With the growing interest in the cognitive capabilities of AI, the discussions on the origins of knowledge and comprehension of reality become increasingly relevant, offering a thought-provoking perspective. Therefore, the following two sections give an overview of the two aforementioned philosophical approaches to later evaluate their applicability to LLMs' mechanisms. However, before delving deeper into this topic, it is important to stress that 'understanding' here solely refers to the process of acquiring knowledge and by no means implies that the discussed systems have gained consciousness.

---

<sup>1</sup> Buckner, C. (2023, March 25). Moderate Empiricism and Machine Learning. The Philosophy of Deep Learning. New York, NY, United States.

<sup>2</sup> Spelke, E. S. (1998). Nativism, empiricism, and the origins of knowledge. *Infant Behavior and Development*, 21(2), 181–200. [https://doi.org/10.1016/s0163-6383\(98\)90002-9](https://doi.org/10.1016/s0163-6383(98)90002-9)

<sup>3</sup> *ibid*

## 2.1. Nativism

Nativism is a philosophical position that asserts the existence of innate or inborn knowledge, suggesting that certain ideas and concepts are present inherently from birth.<sup>4</sup> In other words, our understanding of the world is shaped by inherent truths, which we subscribe to from infancy. This also concerns human abilities and developmental processes that are predetermined and not acquired through experiences.

Plato, the first advocate of nativism, rejected empiricism and argued that our senses cannot be fully trusted as they limit our knowledge to mere opinions derived from sensory experiences. He believed in the existence of innate knowledge through the concept of “recollection”.<sup>5</sup> Plato proposed that humans possess prior knowledge of the Forms, perfect representations of abstract concepts.<sup>6</sup> Through a process of recollection, or simply learning, individuals can delve into this innate knowledge and gain a deeper understanding of the world by revisiting these inherent truths. Similar view was held by René Descartes, influential 17th century nativist philosopher. Descartes' nativism suggests that the soul possesses innate knowledge and experiences before being united with the body.<sup>7</sup> According to Descartes, innate ideas are ‘planted’ in one’s mind by God, serving as a foundational basis for acquiring knowledge.<sup>8</sup>

Despite being a popular view among the philosophers, for a long time nativism came to be seen as unscientific and inferior in comparison to other theories on the origins of knowledge.<sup>9</sup> This only changed in the 1960s with Noam Chomsky and his work in the domain of linguistics. Chomsky argued that children acquire language rapidly and effortlessly, despite the complexities of grammar (Sampson, 2022; BBC Radio 4, 2015). He rejected the notion that this ability can be solely attributed to external factors, such as environment or reinforcement. Instead, Chomsky proposed the existence of an innate language faculty that enables children to acquire language in a relatively efficient manner, independent of specific linguistic environments. This is what he calls “language acquisition device” that humans are born with (Sampson, 2022).

## 2.2. Empiricism

Empiricism thinkers have rejected the idea of innate knowledge, claiming that people are born with *tabula rasa*, blank slate.<sup>10</sup> Empiricists' central argument posits that our knowledge in a subject is acquired through our experiences, whether sensory or reflective. This view was strongly promoted by John Locke, one of the most famous philosophers and political theorists of the 17th century. Despite advocating for the ideas of empiricism, Locke agreed that our senses cannot always be trusted and might be an unreliable source of knowledge.<sup>11</sup> Therefore, he proposed a distinction between primary and secondary qualities. Primary qualities refer to physical qualities of an object, its inherent characteristics that exist regardless of our perception. That includes attributes such as occupying physical space, being in a state of motion or rest, and

---

<sup>4</sup> Markie, P., & Folescu, M. (2023, September 2). Rationalism vs. empiricism. Stanford Encyclopedia of Philosophy. <https://plato.stanford.edu/archives/spr2023/entries/rationalism-empiricism/>

<sup>5</sup> Kraut, R. (2022, February 12). Plato. Stanford Encyclopedia of Philosophy. <https://plato.stanford.edu/archives/spr2022/entries/plato/>

<sup>6</sup> *ibid.*

<sup>7</sup> Hatfield, G. (2018, January 16). René Descartes. Stanford Encyclopedia of Philosophy. <https://plato.stanford.edu/archives/sum2018/entries/descartes/>

<sup>8</sup> UKEssays. (November 2018). A Comparison of Descartes' Nativism and Locke's Empiricism. Retrieved from <https://www.ukessays.com/essays/philosophy/descartes-nativism-vs-locke-empiricism-philosophy-essay.php?vref=1>

<sup>9</sup> Samet, J. (2023, March 27). The historical controversies surrounding innateness. Stanford Encyclopedia of Philosophy. <https://plato.stanford.edu/archives/sum2023/entries/innateness-history/>

<sup>10</sup> Markie, P., & Folescu, M. (2023, September 2). Rationalism vs. empiricism. Stanford Encyclopedia of Philosophy. <https://plato.stanford.edu/archives/spr2023/entries/rationalism-empiricism/>

<sup>11</sup> CrashCourse. (2016, March 15). Locke, Berkeley, & Empiricism: Crash Course Philosophy #6 [Video]. YouTube. <https://www.youtube.com/watch?v=5C-s4JrymKM>

possessing solidity and texture. Secondary qualities exist in our minds and result through the interaction between our senses with the primary qualities of an object.<sup>12</sup>

Another influential author that subscribed to the belief that our understanding is a result of experiences was David Hume. He held the view that knowledge is acquired solely through sensations, emotions, and passions.<sup>13</sup> He simplified the contents of the mind to the concept of perception, which he classified into two categories: impressions and ideas.<sup>14</sup> Impressions encompass the direct sensory input obtained from the senses, passions, and emotions. Ideas, on the other hand, are representations or dim reflections of impressions shaped during the processes of thinking and reasoning.<sup>15</sup> Although Hume is often characterized as anti-nativist, his stance on the innateness debate is more nuanced, as he contends that it is fundamentally misguided. In his account, the origin of ideas holds little philosophical significance. What matters to him is not whether ideas are innate or acquired, but rather their nature and content. The crucial factor is that the resulting idea accurately represents a sensory state derived from actual experience.<sup>16</sup> In other words, the emphasis is again on construing our experiences.

### 3. How do the accounts of nativism and empiricism relate to LLMs?

The two accounts of the origin of knowledge provide a thought-provoking framework for analyzing the advancements made in LLMs, as well as other AI systems. Cameron Buckner in his article *Deep learning: A philosophical introduction*<sup>17</sup> evokes the example of AlphaGo Zero, a computer program that was able to defeat a Go world champion. Using this example, he links the mechanisms of deep neural networks to the accounts of empiricism and nativism. On one hand, preset weights, its algorithm and prebuilt Go rules, oppose the *tabula rasa* component of empiricism and indicate some innate knowledge that this system possesses. On the other hand, the fact that AlphaGo learns its strategy and improves it through self-play, supports the empiricist approach. In the same vein, what distinguishes the newest version of GPT from the earlier models is the implementation of Reinforcement Learning from Human Feedback (RLHF). This technique involves training the model with feedback from human evaluators who act as reward signals to assess the quality of the generated text.<sup>18</sup> The real-world feedback from human evaluators in training language models like GPT-4 aligns with the empiricist approach. This feedback serves as empirical input, a guideline to improve the model's performance, emphasizing the role of experience and observation in acquiring knowledge. On the other hand, some might argue that some information is predetermined in language models. For example, training data is labelled, and dataset are given different weights. This annotation allows LLMs to develop a broad comprehension of text generation.<sup>19</sup>

Whilst it is tempting to suggest a link between nativism and some preset parameters in LLMs, Steven T. Piantadosi<sup>20</sup> believes that models like ChatGPT significantly challenge traditional approaches in linguistics, at the same time undermining Noam Chomsky's claims about innateness in our language. Since these systems can generate human-like prose without

<sup>12</sup> Markie, P., & Folescu, M. (2023, September 2). Rationalism vs. empiricism. Stanford Encyclopedia of Philosophy. <https://plato.stanford.edu/archives/spr2023/entries/rationalism-empiricism/>

<sup>13</sup> Iwuagwu, E. K., & Agabi, G. A. (2019). David Hume's notion of perception and his problem with causality. AFRREV IJAH: An International Journal of Arts and Humanities, 8(4), 66–76. <https://doi.org/10.4314/ijah.v8i4.6>

<sup>14</sup> *ibid*

<sup>15</sup> *ibid*

<sup>16</sup> *Ibid*

<sup>17</sup> Buckner, C. (2019). *Deep learning: A philosophical introduction*. *Philosophy Compass*, 14(10). <https://doi.org/10.1111/phc3.12625>

<sup>18</sup> Malhotra, T. (2023, March 21). Exploring The Differences Between ChatGPT/GPT-4 and Traditional Language Models: The Impact of Reinforcement Learning from Human Feedback (RLHF). MarkTechPost. <https://www.marktechpost.com/2023/03/21/exploring-the-differences-between-chatgpt-gpt-4-and-traditional-language-models-the-impact-of-reinforcement-learning-from-human-feedback-rlhf/>

<sup>19</sup> Kniazieva, Y. (2023, February 23). From Data to Dialogue: Data Annotation for Training AI Chatbots like ChatGPT. Label Your Data. <https://labelyourdata.com/articles/data-annotation-for-training-chatgpt>

<sup>20</sup> Piantadosi, S. T. (2023). Modern language models refute chomsky's approach to language. Lingbuzz Preprint, lingbuzz/007180.

explicit grammatical instruction, showing language is acquired rather than innate.<sup>21</sup> The models' success in various tasks challenges the idea of universal grammar and indicates language can be improved through machine learning. Additionally, the fact that they can produce coherent language on various subjects implies that language is probabilistic and context-dependent rather than rigidly rule-based.<sup>22</sup> Despite Chomsky's substantial contributions, these models provide fresh perspectives on the nature of language and its interaction with machine learning.

Furthermore, LMMs rely on extensive training data to learn patterns and generate responses. However, the models do not possess sensory experiences in the same way humans do. They do not perceive the world through senses but rather analyze textual data. This suggests that these systems cannot replicate sensory experiences that inform human understanding. However, by training Vision-and-Language (VL) models and using text and image (jointly) or video data, Yun et al.<sup>23</sup> highlighted that adding sensory grounding does not improve the performance of these models. This research was a response to common criticism that linguistic representations in these AI systems lack the association with the real-world meaning of words. Yun et al. argue that the structure of representations that the models learn from the language itself resembles what we mean by sensory grounding.<sup>24</sup> The knowledge ingrained in the language allows for acquiring the concepts like, for example, 'north' or 'left'.<sup>25</sup>

#### 4. Conclusion

Our understanding of the human mind and the nature of knowledge is still captivated and challenged by the ongoing philosophical debate between nativism and empiricism. The empiricist position, which emphasizes the importance of experience and observation in learning, contrasts with the nativist perspective, which asserts the existence of intrinsic knowledge. LLMs in their current state can be discussed through the lenses of both, but none of the approaches is applicable to their full extent. The relevance of this debate is reflected in the growing literature that attempts to create an appropriate framework that would capture the intricacies of artificial intelligence. For instance, the theory of connectionism contends that the performance of artificial neural networks improves with reassessing the patterns and relationships by changing connection weight values and more data.<sup>26</sup> On the other hand, scholars like Gary Marcus have advocated for increased emphasis on innateness in artificial intelligence, endorsing the symbolic approach that involves inputting logical rules into our models and creating explicit behavior rules.<sup>27</sup> Comparing LLMs to human understanding is not for us to claim they are human-competitive, but rather to examine how we intend to develop our AI.

There appears to be a consensus that philosophical questions that once puzzled ancient Greek philosophers can now be examined through empirical methods and tested. However, the approaches of nativism and empiricism are not just tools that we can blindly apply in this discussion. In fact, they are greatly challenged by the recent technological advancements. For example, the current state of language models questions established linguistic theories and provides fresh perspectives on language acquisition and probabilistic language emergence.<sup>28</sup>

---

<sup>21</sup> *ibid*

<sup>22</sup> *ibid*

<sup>23</sup> Yun, T., Sun, C., & Pavlick, E. (2021). Does vision-and-language pretraining improve lexical grounding? Findings of the Association for Computational Linguistics: EMNLP 2021. <https://doi.org/10.18653/v1/2021.findings-emnlp.370>

<sup>24</sup> *ibid*

<sup>25</sup> What's the next word in large language models? (2023). *Nature Machine Intelligence*, 5(4), 331–332.

<https://doi.org/10.1038/s42256-023-00655-z>

<sup>26</sup> Jeevanandam, N. (2022). The significance of connectionism in ai. *INDIAai*. <https://indiaai.gov.in/article/the-significance-of-connectionism-in-ai>

<sup>27</sup> Marcus, G. (2018). Innateness, alphazero, and artificial intelligence. *arXiv preprint arXiv:1801.05667*; Dickson, B. (2019, November 17). What is symbolic artificial intelligence?. *TechTalks*. <https://bdtechtalks.com/2019/11/18/what-is-symbolic-artificial-intelligence/>.

<sup>28</sup> Piantadosi, S. T. (2023). Modern language models refute chomsky's approach to language. *Lingbuzz Preprint*, lingbuzz/007180.



Additionally, while these models lack sensory experiences, researchers have shown that the models' ability to learn patterns and words' representation rejects the need for these sensations.<sup>29</sup>

Furthermore, examining AI's cognitive capabilities through the lens of human understanding can shed light on its limitations. For instance, concepts like passions and emotions, central to David Hume's philosophy, cannot be directly applied in the context of LLMs, which are probabilistic models. It's important to clarify that this paper does not imply that these systems possess the depth of understanding found in humans; rather, it aims to highlight avenues for contemplating their architecture. This perspective aligns with Cameron Buckner's<sup>30</sup> call for collaboration between engineers and philosophers, emphasizing the potential to translate philosophical inquiries into practical testing scenarios led by engineers.

## References

1. BBC Radio 4. (2015, January 22). Noam Chomsky on Language Acquisition [Video]. YouTube. <https://www.youtube.com/watch?v=7Cgpfw4z8cw>
2. Buckner, C. (2019). Deep learning: A philosophical introduction. *Philosophy Compass*, 14(10). <https://doi.org/10.1111/phc3.12625>
3. Buckner, C. (2023, March 25). Moderate Empiricism and Machine Learning. *The Philosophy of Deep Learning*. New York, NY, United States.
4. Chadha, A. S. (2023, January 2). *What exactly is chat GPT & How Does It Work? for a 5-year-old!*. Medium. <https://becominghuman.ai/what-exactly-is-chat-gpt-how-does-it-work-for-a-5-year-old-aeccabdfa990>
5. CrashCourse. (2016, March 15). Locke, Berkeley, & Empiricism: Crash Course Philosophy #6 [Video]. YouTube. <https://www.youtube.com/watch?v=5C-s4JrymKM>
6. Dickson, B. (2019, November 17). *What is symbolic artificial intelligence?*. TechTalks. <https://bdtechtalks.com/2019/11/18/what-is-symbolic-artificial-intelligence/>
7. Future of Life Institute. (2023, June 8). Pause Giant AI Experiments: An Open Letter - Future of Life Institute. <https://futureoflife.org/open-letter/pause-giant-ai-experiments/>
8. Guardian News and Media. (2022, July 23). *Google fires software engineer who claims AI chatbot is sentient*. The Guardian. <https://www.theguardian.com/technology/2022/jul/23/google-fires-software-engineer-who-claims-ai-chatbot-is-sentient>
9. Hatfield, G. (2018, January 16). *René Descartes*. Stanford Encyclopedia of Philosophy. <https://plato.stanford.edu/archives/sum2018/entries/descartes/>
10. Hu, K. (2023, February 2). *CHATGPT sets record for fastest-growing user base - analyst note*. Reuters. <https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/>
11. Iwuagwu, E. K., & Agabi, G. A. (2019). David Hume's notion of perception and his problem with causality. *AFRREV IJAH: An International Journal of Arts and Humanities*, 8(4), 66–76. <https://doi.org/10.4314/ijah.v8i4.6>
12. Jeevanandam, N. (2022). *The significance of connectionism in ai*. INDIAai. <https://indiaai.gov.in/article/the-significance-of-connectionism-in-ai>
13. Kniazieva, Y. (2023, February 23). *From Data to Dialogue: Data Annotation for Training AI Chatbots like ChatGPT*. Label Your Data. <https://labeledyourdata.com/articles/data-annotation-for-training-chatgpt>
14. Kraut, R. (2022, February 12). *Plato*. Stanford Encyclopedia of Philosophy. <https://plato.stanford.edu/archives/spr2022/entries/plato/>
15. Malhotra, T. (2023, March 21). *Exploring The Differences Between ChatGPT/GPT-4 and Traditional Language Models: The Impact of Reinforcement Learning from Human Feedback (RLHF)*. MarkTechPost.

<sup>29</sup> Yun, T., Sun, C., & Pavlick, E. (2021). Does vision-and-language pretraining improve lexical grounding? Findings of the Association for Computational Linguistics: EMNLP 2021. <https://doi.org/10.18653/v1/2021.findings-emnlp.370>

<sup>30</sup> Buckner, C. (2023, March 25). Moderate Empiricism and Machine Learning. *The Philosophy of Deep Learning*. New York, NY, United States.

- <https://www.marktechpost.com/2023/03/21/exploring-the-differences-between-chatgpt-gpt-4-and-traditional-language-models-the-impact-of-reinforcement-learning-from-human-feedback-rlhf/>
16. Markie, P., & Folescu, M. (2023, September 2). *Rationalism vs. empiricism*. Stanford Encyclopedia of Philosophy. <https://plato.stanford.edu/archives/spr2023/entries/rationalism-empiricism/>
  17. Marcus, G. (2018). Innateness, alphazero, and artificial intelligence. arXiv preprint arXiv:1801.05667.
  18. Piantadosi, S. T. (2023). Modern language models refute chomsky's approach to language. Lingbuzz Preprint, ling- buzz/007180.
  19. Ruby, M. (2023, February 16). *How CHATGPT works: The models behind the bot*. Medium. <https://towardsdatascience.com/how-chatgpt-works-the-models-behind-the-bot-1ce5fca96286>
  20. Samet, J. (2023, March 27). *The historical controversies surrounding innateness*. Stanford Encyclopedia of Philosophy. <https://plato.stanford.edu/archives/sum2023/entries/innateness-history/>
  21. Sampson, G. (2022). *Nature or Nurture?*. Empiricism v. Nativism. <https://www.grsampson.net/REmpNat.html>
  22. Spelke, E. S. (1998). Nativism, empiricism, and the origins of knowledge. *Infant Behavior and Development*, 21(2), 181–200. [https://doi.org/10.1016/s0163-6383\(98\)90002-9](https://doi.org/10.1016/s0163-6383(98)90002-9)
  23. The Philosophy of Deep Learning. (n.d.). <https://phildeeplearning.github.io/#livestream>
  24. UKEssays. (November 2018). A Comparison of Descartes' Nativism and Locke's Empiricism. Retrieved from <https://www.ukessays.com/essays/philosophy/descartes-nativism-vs-locke-empiricism-philosophy-essay.php?vref=1>
  25. What's the next word in large language models? (2023). *Nature Machine Intelligence*, 5(4), 331–332. <https://doi.org/10.1038/s42256-023-00655-z>
  26. Yun, T., Sun, C., & Pavlick, E. (2021). Does vision-and-language pretraining improve lexical grounding? *Findings of the Association for Computational Linguistics: EMNLP 2021*. <https://doi.org/10.18653/v1/2021.findings-emnlp.370>

## SHOULD WE CALL THE HALLUCINATING ORACLE AN EXPERT?: LARGE LANGUAGE MODELS AND THE CONCEPT OF EXPERTISE

THOMAS WACHTER

*Utrecht University, Utrecht, The Netherlands*

*E-mail: [t.g.wachterwielandt@students.uu.nl](mailto:t.g.wachterwielandt@students.uu.nl)*

This paper delves into the intricate relationship between expertise and large language models (LLMs), such as ChatGPT. Drawing on a philosophical exploration of expertise, it addresses the fundamental question of whether LLMs can be considered experts. The author navigates contrasting perspectives in the field, with some arguing that understanding is a crucial component of expertise, while others emphasize competence and results. The paper underscores the significant challenge posed by LLMs' ability to generate text that appears to exhibit understanding, while in reality, it relies on statistical patterns. It raises concerns about the potential consequences of regarding LLMs as experts, particularly in contexts where facts are crucial and their propensity for "hallucinating" information. The paper ultimately calls for a reevaluation of existing theories of expertise in light of these technologically advanced models, emphasizing the need for greater clarity and public understanding of their capabilities.

### 1. Introduction

Because “no individual is self-sufficient”<sup>1</sup> we usually rely on and depend on others when we acquire knowledge. When this relationship is asymmetric, and the other knows more than us, we call it an “expert.” In that sense, “expert” is a contrastive term to denominate someone more competent in the domain at issue.<sup>2</sup> Moreover, Grundmann says, experts typically occupy powerful and highly influential social roles, influencing individual laypeople, public opinion, and political deliberation.<sup>3</sup> This epistemic dependence<sup>4</sup> happens everywhere and every time and is especially evident in societies with high levels of specialization.<sup>5</sup> A vast amount of philosophy has been done to capture what we understand by the term “expert”<sup>1</sup> with no clear solution and some had even argued that there might be a different criterion in different contexts.<sup>6</sup> In his paper, Scholz, trying to characterize experts, points out different “symptoms” of expertise.<sup>7</sup> One of the most problematic of these symptoms is “understanding.” Some authors defend understanding-linked accounts of expertise, arguing that either is an important characteristic of experts.<sup>8</sup> In contrast, others think that asking for understanding requires too much of experts.<sup>9</sup>

Large language models (LLMs) like ChatGPT add a new layer of complexity to the study of expertise and challenge our current theories of knowledge, as Goldman predicted at the beginning of the century.<sup>10</sup> The question of what constitutes an expert in relation to AI becomes

---

<sup>1</sup> Plato (1941). *The Republic*. Oxford University Press.

<sup>2</sup> Scholz, O. R. (2018). Symptoms of expertise: Knowledge, understanding and other cognitive goods. *Topoi*, 37:29–37.

<sup>3</sup> Grundmann, T. (2022). Experts: What are they and how can laypeople identify them? In Lackey, J. and McGlynn, A., editors, *Oxford Handbook of Social Epistemology*. Oxford University Press.

<sup>4</sup> Hardwig, J. (1985). Epistemic dependence. *The Journal of Philosophy*, 82(7):335–349.

<sup>5</sup> Scholz, O. R. (2018). Symptoms of expertise: Knowledge, understanding and other cognitive goods. *Topoi*, 37:29–37.

<sup>6</sup> Goldman, A. (2018). Expertise. *Topoi*, 37:3–10.

<sup>7</sup> Scholz, O. R. (2018). Symptoms of expertise: Knowledge, understanding and other cognitive goods. *Topoi*, 37:29–37.

<sup>8</sup> Croce, M. (2019). On what it takes to be an expert. *The Philosophical Quarterly*, 69, Scholz, O. R. (2018). Symptoms of expertise: Knowledge, understanding and other cognitive goods. *Topoi*, 37:29–37.

<sup>9</sup> Grundmann, T. (2022). Experts: What are they and how can laypeople identify them? In Lackey, J. and McGlynn, A., editors, *Oxford Handbook of Social Epistemology*. Oxford University Press.

<sup>10</sup> Goldman, A. I. (2000). Telerobotic knowledge: A reliabilist approach. In Goldberg, K., editor, *The Robot in the Garden*, pages 126–142. MIT Press, Cambridge, MA.

very relevant when millions of people are using these technologies,<sup>11</sup> and a rising number of products are built on top of these models. Before deploying these systems at a large scale, we should answer many ethical and philosophical questions.

Through this examination, I will argue that calling these models experts and, perhaps more importantly, interacting with them as such could be dangerous for truth and democracy, especially when these models “hallucinate”<sup>12</sup> but, at the same time, our current theoretical frameworks for evaluating these models doesn’t help much. Therefore, in this paper, I will examine various theories of expertise, emphasizing the discussion on the role of understanding and showing that LLMs call for re-evaluating our understanding of expertise.

## 2. Large Language Models

LLMs are systems trained, based on an extensive corpus of documents, to be able to predict the next token (word, character, or string).<sup>13</sup> The training process results in a complex statistical model of how the words and phrases relate.<sup>14</sup> With the advancements in deep learning (DL), the increasing amount of data, new ways of representing the distribution of words in the text (word embedding<sup>3</sup>), and the introduction of “transformers”<sup>15</sup>, the field of natural language processing (NLP) has seen tremendous advances. Nowadays, we can freely use models like GPT-4<sup>16</sup> that “can produce astonishingly human-like text, conversation, and, in some cases, what seems like human reasoning abilities”.<sup>17</sup> These amazing results make people wonder about what these models are.

## 3. Understanding and Hallucinations

The debate if language models “understand” dates back to AI’s early conceptions. Even though some claimed that those early developments had all kinds of mental capabilities,<sup>18</sup> there was an implicit agreement that “while AI systems exhibit seemingly intelligent behaviour in many specific tasks, they do not understand the data they process in the way humans do”.<sup>19</sup> However, with the development of LLMs, people have changed their views about machines and understanding. A 2022 survey given to active NLP researchers shows how divided the field is concerning this topic. One survey item asked if the respondent agreed with the following statement about whether LLMs could ever, in principle, understand language. Of 480 people responding, essentially half (51%) agreed, and the other half (49%) disagreed.<sup>20</sup> On the one hand, there exist those who argue that however fluent in their linguistic output, LLMs are “stochastic parrots”<sup>21</sup> that “cannot possess understanding because they have no experience or mental models of the world”<sup>22</sup>. On the other, some argue that LLMs are constructing a new type

<sup>11</sup> Buchholz, K. (2023). Time to one million users. Statista. Published on January 24, 2023.

<sup>12</sup> OpenAI (2023). Gpt-4 technical report ; Bang, Y., Cahyawijaya, S., Lee, N., Dai, W., Su, D., Wilie, B., Lovenia, H., Ji, Z., Yu, T., Chung, W., Do, Q. V., Xu, Y., and Fung, P. (2023). A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity.

<sup>13</sup> Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? . In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, FAccT ’21, page 610–623, New York, NY, USA. Association for Computing Machinery.

<sup>14</sup> Mitchell, M. and Krakauer, D. C. (2022). The debate over understanding in ai’s large language models.

<sup>15</sup> Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, , and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30:5998–6008.

<sup>16</sup> OpenAI (2023). Gpt-4 technical report

<sup>17</sup> Mitchell, M. and Krakauer, D. C. (2022). The debate over understanding in ai’s large language models.

<sup>18</sup> Mitchell, M. (2021). Why ai is harder than we think. In Proceedings of the Genetic and Evolutionary Computation Conference, GECCO ’21, page 3, New York, NY, USA. Association for Computing Machinery.

<sup>19</sup> Mitchell, M. and Krakauer, D. C. (2022). The debate over understanding in ai’s large language models.

<sup>20</sup> Michael, J., Holtzman, A., Parrish, A., Mueller, A., Wang, A., Chen, A., Madaan, D., Nangia, N., Pang, R. Y., Phang, J., and Bowman, S. R. (2022). What do nlp researchers believe? results of the nlp community metasurvey.

<sup>21</sup> Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? . In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, FAccT ’21, page 610–623, New York, NY, USA. Association for Computing Machinery.

<sup>22</sup> Mitchell, M. and Krakauer, D. C. (2022). The debate over understanding in ai’s large language models.

of understanding based on the exploitation of unforeseen statistical relations between words<sup>23</sup> or even that they are decoupling human understanding.<sup>24</sup>

To do philosophy on this, I have to clarify some things, especially what LLMs really do. In a recent article, Shanahan gives a great example of what LLMs do: suppose a give a prompt to an LLM of the type “The first person to walk on the Moon was, ” and it responds “Neil Armstrong.” Why is that? Shanahan points out that “in an important sense, we are not asking who was the first person to walk on the Moon. We are asking the model the following question: Given the statistical distribution of words in the vast public corpus of (English) text, what words are most likely to follow the sequence “The first person to walk on the Moon was ?” A good reply to this question is “Neil Armstrong.”<sup>25</sup> To the human user, the moon represents a real truth. It is tied to reality.<sup>26</sup> In other words, these models are not trained for truth. They just predict the next token. This distinction is fundamental and can be risky if overlooked. Individuals must acknowledge that these models should not be treated as mere search engines or oracles capable of accurately predicting the future. A notable instance illustrating this concern is the well-documented case of a lawyer in New York who relied on ChatGPT for a legal matter, only to discover that the model generated fictional cases, leading to misleading information.<sup>27</sup> To further illustrate, I engaged in a conversation with ChatGPT, in which I sought factual information about the world:

USER: Can you give me some literature recommendations for a philosophy paper that studies the epistemic status of large language models?  
 CHATGPT: 1. “Epistemic challenges of machine learning” by Zachary C. Lipton.  
 2. “Epistemic opacity in neural networks” by Jesse Mu and Benjamin Van Roy.  
 3. “The Knowledge Illusion: Why We Never Think Alone” by Steven Sloman and Philip Fernbach.

Figure 1: Dialogue between ChatGPT and me (18/06/2023)

After searching, I realized that even though the scholars’ names are real names, only “The Knowledge Illusion” is a real source. The example in figure 1 shows the phenomena known as “hallucinations”.<sup>28</sup> This makes these models unreliable, especially when facts are important.<sup>29</sup> In addition, as also mentioned in the OpenAI report, as these models are integrated into society, this tendency can lead to the degradation of information and reduce trust,<sup>30</sup> which is one of the pillars of our society. But, does it mean that we should not consider them experts?

Not considering the problem of understanding, and just following the route of competence, Shannahan argues that some problems can be broken down into multiple inference steps, and

<sup>23</sup> Mitchell, M. and Krakauer, D. C. (2022). The debate over understanding in ai’s large language models; Strasser, A. (2023). A new kind of comprehension in large language models. In Hybrid Workshop on the Philosophy of Large Language Models, Eindhoven, Netherlands.

<sup>24</sup> Mahowald, K., Ivanova, A. A., Blank, I. A., Kanwisher, N., Tenenbaum, J. B., and Fedorenko, E. (2023). Dissociating language and thought in large language models: a cognitive perspective.

<sup>25</sup> Shanahan, M. (2022). Talking about large language models.

<sup>26</sup> *ibid*

<sup>27</sup> Weiser, B. and Schweber, N. (2023). The chatgpt lawyer explains himself. New York Times.

<sup>28</sup> Bang, Y., Cahyawijaya, S., Lee, N., Dai, W., Su, D., Wilie, B., Lovenia, H., Ji, Z., Yu, T., Chung, W., Do, Q. V., Xu, Y., and Fung, P. (2023). A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity ; OpenAI (2023). Gpt-4 technical report.

<sup>29</sup> OpenAI (2023). Gpt-4 technical report.

<sup>30</sup> OpenAI (2023). Gpt-4 technical report.

external sources can be consulted in each step.<sup>31</sup> Recent articles have shown advancements in this area,<sup>32</sup> revealing that while employing multi-step reasoning, the model exhibits improved abilities in inducing, tracking, and updating action plans and effectively handling exceptions. Furthermore, the model can interface with external sources, such as knowledge bases or environments, allowing it to gather additional information.<sup>33</sup> This might suggest that the problems of Hallucinations are solvable in the future and might not depend on understanding. This plausible solution is interesting and worth pursuing, but it raises the question of whose “truth” we are going to portray, but this goes out of the scope of this paper. In the next section, I will discuss understanding in the context of epistemic expertise.

### 3.1. Understanding in epistemic expertise

Understanding is a complicated concept. Despite the enthusiasm, it remains faintly understood.<sup>34</sup> Even more importantly, the relationship between understanding and expertise must be clarified. In epistemology, philosophers still do not agree on the role that understanding plays in expertise. On the one hand, scholars like Grundmann and Goldman defend a view of expertise that Croce defines as novice-orientated approach.<sup>35</sup> These functionalist definitions are based on what experts can do. On the other hand, Croce<sup>36</sup> and Scholz<sup>37</sup> emphasize experts’ characteristics, arguing especially in favor of mental attributes like understanding. In the case of the functionalist approach, Grundmann says that someone should be considered an expert in a particular domain *D* at time *t* if (i) they must have access to more information and facts at *t*, (ii) be able to process that information more reliably and efficiently than most others at *t*, and (iii) be considered proficient or master in the domain based on the two previous criteria at *t*.<sup>38</sup> Regarding understanding, it is clear why this definition does not require it. Grundmann argues that requiring understanding is too demanding for experts.<sup>39</sup> If we apply Grundmann’s definition of expertise to LLMs we can see that they satisfy the definition. For (i) LLMs have more evidence than every human being. These models are trained with huge amounts of data collected from the internet.<sup>40</sup> For (ii) is not that clear; reasoning in LLMs is hard to settle because reasoning is content-neutral, but even though the complexity of the term reasoning “LLMs can be effectively applied to multi-step reasoning without further training, thanks to clever prompt engineering.”<sup>41</sup> Therefore, following Grundmann’s definition, we should consider LLMs as epistemic experts. On the other side, Croce and Scholz defend an understanding-linked account, emphasizing the value of understanding for epistemic expertise. For example, Scholz argues that “*E* is an expert in domain *D* if: *E* has a considerably better understanding of domain *d* than the vast majority of people do”.<sup>42</sup> He concedes that understanding is not a necessary condition but at least an important symptom of expertise.<sup>43</sup> In the same line of argumentation, Croce argues that a subject *S* can be considered an expert if and only if it possesses a better understanding of

<sup>31</sup> Shanahan, M. (2022). Talking about large language models.

<sup>32</sup> Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., and Cao, Y. (2023). React: Synergizing reasoning and acting in language models

<sup>33</sup> *ibid*

<sup>34</sup> Khalifa, K. (2013). Is understanding explanatory or objectual? *Synthese*, 190(6):1153–1171.

<sup>35</sup> Croce, M. (2019). On what it takes to be an expert. *The Philosophical Quarterly*, 69:1.

<sup>36</sup> *ibid*

<sup>37</sup> Scholz, O. R. (2018). Symptoms of expertise: Knowledge, understanding and other cognitive goods. *Topoi*, 37:29–37.

<sup>38</sup> Grundmann, T. (2022). Experts: What are they and how can laypeople identify them? In Lackey, J. and McGlynn, A., editors, *Oxford Handbook of Social Epistemology*. Oxford University Press.

<sup>39</sup> *ibid*

<sup>40</sup> Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? . In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, FAccT ’21*, page 610–623, New York, NY, USA. Association for Computing Machinery.

<sup>41</sup> Shanahan, M. (2022). Talking about large language models.

<sup>42</sup> Scholz, O. R. (2018). Symptoms of expertise: Knowledge, understanding and other cognitive goods. *Topoi*, 37:29–37.

<sup>43</sup> *ibid*

D than most people do.<sup>44</sup> In addition, he says that we should consider “understanding as a fundamental epistemic goal ... experts are supposed to understand the relationships between the various components of D.”<sup>45</sup> We could follow a consequentialist approach here and argue that if LLMs produce good results and cite their sources, we should not care about understanding. But on the other hand, arguably, as shown before, the lack of understanding of LLMs is one of the main reasons for their failure. These models are not embodied beings that triangulate with the world in order to update their beliefs,<sup>46</sup> and therefore can “hallucinate” facts with high confidence and coherence. The problem that we encounter is summarized in the following table:

	Evidence-linked	Understanding-linked
LLMs-understand	Experts	Experts
LLMs-Don't understand	Experts	No Experts

One question that arises from this discussion is, what do we want? Do we want to introduce these types of experts, or do we not? I believe this is an important question to tackle as a society that we have not answered yet, which is worrying because we are using these systems without knowing how to categorize them. In addition, and interestingly enough, LLMs seem to serve as examples in both cases. Meaning that, on the one hand, evidence-linked supporters of epistemic expertise, like Grundmann, could use them to argue that they are able to produce meaningful outputs without understanding, so understanding is not important. On the other hand, understanding-linked defenders like Scholz could argue that the lack of understanding of these models is the key factor that produces their failures, therefore, understanding is necessary. This shows how our current theories of knowledge must be updated, as Grundmann predicted, due to the rise of this technology. Apart from the technical solutions, I see a future in the idea of building new concepts for technology-based expertise<sup>47</sup> while informing the public about what actually these models do.

#### 4. Conclusions

In conclusion, I showed the two different discussions that collide when we try to categorize LLMs as epistemic expertise. On the one hand, evidence-linked accounts, propose that understanding is not a necessary feature of expertise, while understanding-linked accounts, consider understanding as an important symptom of expertise. On the other hand, the discussion on understanding in the context of LLMs is increasing due to the impressive results of models such as ChatGPT. Some argue that these models have no understanding of the world and they learn the statistical relationships between words. In contrast, others argue that these models have a new type of understanding that should also be considered valid. Finally, more than trying to give a clear answer to the question if LLMs should be considered epistemic experts or not, I showed how its development adds a new layer of complexity to the study of experts and that calling them experts or interacting with them as such is dangerous at least in current states. This is an essential topic for today's society, and this article tries to add a little bit of clarity to this vast but important topic.

#### References

1. Bang, Y., Cahyawijaya, S., Lee, N., Dai, W., Su, D., Wilie, B., Lovenia, H., Ji, Z., Yu, T., Chung, W., Do, Q. V., Xu, Y., and Fung, P. (2023). A multitask, multilingual, multimodal

<sup>44</sup> Croce, M. (2019). On what it takes to be an expert. *The Philosophical Quarterly*, 69:1.

<sup>45</sup> *ibid*

<sup>46</sup> Shanahan, M. (2022). Talking about large language models.

<sup>47</sup> Freiman, O. (2023). Analysis of beliefs acquired from a conversational ai: Instruments-based beliefs, testimony-based beliefs, and technology-based beliefs. *Episteme*, pages 1–17.

- evaluation of chatgpt on reasoning, hallucination, and interactivity.
2. Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? . In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, FAccT '21*, page 610–623, New York, NY, USA. Association for Computing Machinery.
  3. Buccholz, K. (2023). Time to one million users. Statista. Published on January 24, 2023.
  4. Croce, M. (2019). On what it takes to be an expert. *The Philosophical Quarterly*, 69:1.
  5. Freiman, O. (2023). Analysis of beliefs acquired from a conversational ai: Instruments-based beliefs, testimony-based beliefs, and technology-based beliefs. *Episteme*, pages 1–17.
  6. Goldman, A. (2018). Expertise. *Topoi*, 37:3–10.
  7. Goldman, A. I. (2000). Telerobotic knowledge: A reliabilist approach. In Goldberg, K., editor, *The Robot in the Garden*, pages 126–142. MIT Press, Cambridge, MA.
  8. Grundmann, T. (2022). Experts: What are they and how can laypeople identify them? In Lackey, J. and McGlynn, A., editors, *Oxford Handbook of Social Epistemology*. Oxford University Press.
  9. Hardwig, J. (1985). Epistemic dependence. *The Journal of Philosophy*, 82(7):335–349.
  10. Khalifa, K. (2013). Is understanding explanatory or objectual? *Synthese*, 190(6):1153–1171.
  11. Mahowald, K., Ivanova, A. A., Blank, I. A., Kanwisher, N., Tenenbaum, J. B., and Fedorenko, E. (2023). Dissociating language and thought in large language models: a cognitive perspective.
  12. Michael, J., Holtzman, A., Parrish, A., Mueller, A., Wang, A., Chen, A., Madaan, D., Nangia, N., Pang, R. Y., Phang, J., and Bowman, S. R. (2022). What do nlp researchers believe? results of the nlp community metasurvey.
  13. Mitchell, M. (2021). Why ai is harder than we think. In *Proceedings of the Genetic and Evolutionary Computation Conference, GECCO '21*, page 3, New York, NY, USA. Association for Computing Machinery.
  14. Mitchell, M. and Krakauer, D. C. (2022). The debate over understanding in ai’s large language models.
  15. OpenAI (2022). Chatgpt: Optimizing language models for dialogue. OpenAI Blog. <https://openai.com/blog/chatgpt/>.
  16. OpenAI (2023). Gpt-4 technical report.
  17. Plato (1941). *The Republic*. Oxford University Press.
  18. Scholz, O. R. (2018). Symptoms of expertise: Knowledge, understanding and other cognitive goods. *Topoi*, 37:29–37.
  19. Shanahan, M. (2022). Talking about large language models.
  20. Strasser, A. (2023). A new kind of comprehension in large language models. In *Hybrid Workshop on the Philosophy of Large Language Models*, Eindhoven, Netherlands.
  21. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, , and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30:5998–6008.
  22. Weiser, B. and Schweber, N. (2023). The chatgpt lawyer explains himself. *New York Times*.
  23. Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., and Cao, Y. (2023). React: Synergizing reasoning and acting in language models



## MAKING HUMAN PROGRESS OR FALLING BEHIND: DOES CHATGPT EXACERBATE THE KNOWLEDGE GAP?

JIAYI LIU

*Utrecht University, Utrecht, The Netherlands*

*E-mail: [j.liu12@students.uu.nl](mailto:j.liu12@students.uu.nl)*

This paper explores the multifaceted impact of ChatGPT, an advanced language model developed by OpenAI, on human knowledge and the digital divide. ChatGPT's impressive capabilities have led to its rapid adoption by a wide range of users, raising questions about its role as a tool for human progress versus a crutch for those unwilling to think. The paper delves into the theoretical frameworks of extended mind theory, where technology supplements cognitive processes, and the knowledge gap/digital divide, which highlights disparities in access to and understanding of technology. It posits that ChatGPT, while enhancing productivity, may also create dependency, potentially hindering essential cognitive skills. Moreover, ChatGPT may exacerbate the knowledge gap by requiring digital literacy, financial resources, and positive attitudes towards technology, thus benefiting those already knowledgeable while leaving others further behind. The paper concludes by emphasizing the need for a balanced approach in harnessing AI's potential while addressing the knowledge gap to ensure equitable access and effective utilization of technology.

### 1. Introduction

ChatGPT is a language model developed by OpenAI, which reached 1 million users only in 5 days, which can be seen as a revolutionary understanding of natural language <sup>1</sup>. This technology signifies the first comprehensive understanding of the world's knowledge structure. Coupled with the ability to communicate with humans, this achievement represents a significant milestone in artificial intelligence. It means language models like this artificial intelligence have progressed from perception to the principles of brain function and cognitive level. ChatGPT's primary characteristic is understanding and responding to input in natural language. It uses a model of natural language proceeding (NLP) to learn user input and output answers.<sup>2</sup> One of the main characteristics that distinguish ChatGPT from other AI is that this model makes ChatGPT more natural and intuitive, more like a person when conversing with it. ChatGPT's abilities include the features like remembering aspects, supportive communication, follow-up corrections, etc.<sup>3</sup> It often serves as a chatbot that imparts particular technical details, which are developed by engineers and then can be accessed via various platforms such as a website, smartphone application, or a messaging service because of its open API. ChatGPT has big potential applications since it was developed, particularly in diverse fields such as art and technology. It is remarkable how different individuals can use it for various purposes, revealing both its versatility and potential areas for improvement. According to personal experience using ChatGPT published on the Internet, individuals can utilize ChatGPT to accomplish nearly 20 distinct tasks in the digital realm, which encompass activities such as creation production, code interpretation, and information extraction, among others. People can ask ChatGPT to explain the essence of certain concepts in an easy-to-understand manner, provide relevant examples, and suggest areas of content to further study based on our current situation. As a researcher, for instance, ChatGPT can serve as an academic assistant, requesting it to organize learning notes,

---

<sup>1</sup> Firat, M. (2023a). How chat GPT can transform autodidactic experiences and open education. *Department of Distance Education, Open Education Faculty, Anadolu Unive*

<sup>2</sup> LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436-444.

<sup>3</sup> Firat, M. (2023b). How chat GPT can transform autodidactic experiences and open education. *Department of Distance Education, Open Education Faculty, Anadolu Unive*

present a few examples supporting certain viewpoints, or even propose contrasting perspectives. And scholars and researchers are making use of ChatGPT to contribute to the further development of the world. It can inspire health workers to propose new solutions to fight disease and social difficulties and do data analysis and interpretation to gain valuable insights into climate trends.<sup>4</sup> Undeniably, the impact of ChatGPT on human development holds revolutionary significance. ChatGPT sets itself apart from previous forms of artificial intelligence because of its genuine capacity to assist people in problem-solving. It allows ordinary users to truly appreciate the help that ChatGPT can provide in any professional or educational context without any barriers to entry. Some people use it as a tool to make their lives more convenient, essentially liberating their hands, while others utilize it to create more complex and sophisticated products. To deeply explore this phenomenon, I plan to use ChatGPT as an example, focusing on the key question: Making human progress or falling behind, does ChatGPT exacerbate the knowledge gap? The question will be divided into two questions: RQ 1: How does ChatGPT serve as a tool to aid human progress rather than a crutch for those unwilling to think? RQ 2: Does ChatGPT exacerbate the knowledge gap?

## 2. Theoretical frameworks

### 2.1. *Extended mind theory*

The concept of the extended mind “involves the idea that the use of various artifacts, or aspects of the environment can facilitate, enhance or even constitute cognition”.<sup>5</sup> According to Clark and Chalmers (1998) proposed, the cognitive process involves different types of technology. When we try to remember, imagine, or think about solutions to problems, technology can serve as a vehicle for cognition, supplementing or replacing neural mechanisms. As a result, our cognitive abilities are expanded. It means that the mind does not entirely exist in the brain or even the body, but extends into the physical world.

The human mind has expanded to use tools as the source of his thinking. There are many instances showing how our cognitive capabilities can be amplified through technology – smartphones, GPS, online search engines, etc.<sup>6</sup> Humans can keep memories and form ideas by using these tools, essentially leveraging these external devices to carry out cognitive functions. Certain technologies and media are used for different objectives to help people achieve goals. It provides opportunities that carry people’s cognitive processes in particular directions. These processes can significantly impact us and the way we think.

Three criteria are included as external physical processes become part of an individual’s cognitive process:

- That the external resource be reliably available.
- That any information should not usually be subject to critical scrutiny. It should be deemed about as trustworthy as something retrieved clearly from biological memory.
- That information contained in the resource should be easily accessible as and when required (Clark, 2008, 79).

### 2.2. *Knowledge Gap and Digital Divide*

The knowledge gap hypothesis proposed by Tichenor et al. (1970)<sup>7</sup> states that the media can increase gaps in knowledge. "As the infusion of mass media information into a social system

---

<sup>4</sup> Biswas, S. S. (2023b). Potential use of chat gpt in global warming. *Annals of Biomedical Engineering*, 51(6), 1126-1127; Biswas, S. S. (2023a). Role of chat gpt in public health. *Annals of Biomedical Engineering*, 51(5), 868-869; Surameery, N. M. S., & Shakor, M. Y. (2023). Use chat gpt to solve programming bugs. *International Journal of Information Technology & Computer Engineering (IJITC)* ISSN: 2455-5290, 3(01), 17-22.

<sup>5</sup> Gallagher, S. (2013). The socially extended mind. *Cognitive Systems Research*, 25-26, 4-12. <http://doi.org/10.1016/j.cogsys.2013.03.008>.

<sup>6</sup> *ibid*

<sup>7</sup> Tichenor, P. J., Donohue, G. A., & Olien, C. N. (1970). Mass media flow and differential growth in knowledge. *Public Opinion Quarterly*, 34(2), 159-170.

increases, segments of the population with higher socioeconomic status tend to acquire this information at a faster rate than the lower status segments so that the gap in knowledge between these segments tends to increase rather than decrease”.<sup>8</sup>

Knowledge gaps are the result of differences in motivations and the ability to process information.<sup>9</sup> Because people with different socioeconomic status and knowledge, they get access to information and understand information in different ways<sup>10</sup> And mass media aggregate this gap<sup>11</sup>.

The concept of the digital divide was defined as a gap between those who have access to digital technologies and those who do not, originating from knowledge gap.<sup>12</sup> It focuses on the influence of information and communication technologies (ICTs). Some researchers have already made evidence that a “usage gap” between those who use digital technologies for information and those who use it largely for entertainment.<sup>13</sup> While the declining costs of computers and internet access appear to democratize information and knowledge, individuals of lower socio-economic status (SES) who lack foundational knowledge often face significant study and financial barriers. This situation risks creating a self-perpetuating cycle.<sup>14</sup> New knowledge and digital technologies place an increased emphasis on independent learning and rapid growth, offering fresh opportunities. However, those lacking a solid knowledge base are at risk of falling further behind their technologically adept peers, widening the existing knowledge gap.

There are four key issues add up the barriers of using technology and address the digital divide, which are physical access to ICTs, ICT skills and support, attitudes, and content.<sup>15</sup> Physical access means lack of a robust telecommunication infrastructure with sufficient reliable bandwidth for Internet connections, and cost, the ability to purchase, rent or travel to utilize without financial hardship, and the necessary equipment. Lack of ICT skills and support is a significant factor, especially for people who lack computing and technology skills. Cultural and behavioral attitudes towards technology is also essential, for example, some technologies are only for “brainy” people concentrated on the male, young and middle class. And some people may consider the safety of technology such as personal information. Another reason some groups don’t access the new technology is that they are not interested in the content.

### **3. How does ChatGPT serve as a tool to aid human progress rather than a crutch for those unwilling to think?**

ChatGPT facilitate more efficient productivity and sharpen people's logical thinking. Individuals can conserve their time previously spent on navigating through extraneous information and addressing tedious, pointless tasks. With many people use ChatGPT to help them study and do tedious things, thus freeing their hands, an emerging trend where humans are increasingly relying on ChatGPT to assist with their daily tasks may lead to new challenges. For example, the risk of making people overly dependent, potentially diminishing their hands-on, writing, and thinking skills. If the technology becomes unavailable someday, people may resist reverting to handling tedious tasks manually and could even struggle with executing simple, repetitive tasks they once performed.

The act of using technology may not directly contribute to the cognitive process, but the interaction with social customs of problem-solving using technology can be considered a

---

<sup>8</sup> Ibid pp. 159 -160.

<sup>9</sup> Weenig, M. W., & Midden, C. J. (1997). Mass-Media Information Campaigns and Knowledge-Gap Effects 1. *Journal of Applied Social Psychology*, 27(11), 945-958.

<sup>10</sup> Gaziano, C. (2010). Notes on “revisiting the knowledge gap hypothesis: A meta-analysis of thirty-five years of research”. *Journalism & Mass Communication Quarterly*, 87(3-4), 615-632.

<sup>11</sup> Weenig, M. W., & Midden, C. J. (1997). Mass-Media Information Campaigns and Knowledge-Gap Effects 1. *Journal of Applied Social Psychology*, 27(11), 945-958.

<sup>12</sup> Wei, L., & Hindman, D. B. (2011). Does the digital divide matter more? Comparing the effects of new media and old media use on the education-based knowledge gap. *Mass Communication and Society*, 14(2), 216-235.

<sup>13</sup> ibid

<sup>14</sup> Persaud, A. (2001). The knowledge gap. *Foreign Affairs*, 107-117.

<sup>15</sup> Cullen, R. (2001). Addressing the digital divide. *Online Information Review*, 25(5), 311-320.

cognitive component and ChatGPT becomes the role of it. And it is only when people engage appropriately with technology, that it contributes to the formation of cognitive processes. As an external environment, ChatGPT can be instantly accessible to people at any time, and automatically gets people's approval, it automatically excludes a part of people's cognitive process and accelerates the effective cognition solution. People call for ChatGPT as a part of the thinking process and a source of memories. It seems like that ChatGPT already acts as part of the mind and can be seen as an extension of oneself.

Correct and involved interaction with mental institutions makes technology an integral aspect of cognitive processes. Just as drivers using GPS systems too often may lead human's hippocampus to shrink,<sup>16</sup> it is reasonable to think that ChatGPT can increase some abilities and decrease certain abilities at the same time when we put the human brain and tools as a whole. AI technology like ChatGPT might rebalance the academic skill set. The loss of certain skills might not necessarily be problematic, while how to use AI like ChatGPT becomes essential.

#### 4. Does ChatGPT exacerbate the knowledge gap?

It is quite plausible that ChatGPT could exacerbate the knowledge gap, or digital divide. Individuals from various socio-economic statuses may have differing understandings of ChatGPT. Given that their knowledge base may be limited, those who lack related knowledge might not be able to fully utilize the capabilities of ChatGPT.

Based on the four key barriers to technology use and the digital divide outlined by Cullen,<sup>17</sup> we can see several ways in which ChatGPT could potentially exacerbate the knowledge gap. First, ChatGPT requires reliable internet access and a device capable of running it. A simple example is the advanced function requires a monthly subscription of 24 dollars after it released the GPT-4. People who haven't enough financial ability or have no willingness to pay thus cannot access and benefit from ChatGPT. GPT-3 could generate writing that closely resembles human language and give answers to people who seek help, while various limitations have also been observed, such as sometimes generation of incorrect information, or make wrong calculations.<sup>18</sup> GPT-4 exhibits a higher degree of reliability, creativity, and subtlety. With the support of the GPT-4 model, ChatGPT is capable of participating in numerous concurrent discussions, comprehending and replying to natural language inputs, and providing personalized and dynamic assistance.<sup>19</sup> Additionally, it incorporated the ability to handle various things like drawings and videos, functioning as a plug-in when integrated with other tools.

Second, to effectively utilize ChatGPT, users need a basic level of digital literacy. For example, before the plugin stores appeared, people who master computer knowledge and develop code can make plugins to ChatGPT, which create more powerful content and tools, and they can even use ChatGPT to optimize their code. while people who do not know can only still use the basic functions of ChatGPT. Many people today still think of ChatGPT as a search engine and some people already developed program with the help of ChatGPT. Additionally, people some individuals may perceive advanced technology like ChatGPT as violating people's privacy or ethical issue, or have the idea of feeling it's useless because it sometimes gives wrong answers, which could deter certain groups from using it. People with different age and class having different acknowledge towards ChatGPT, thus furthering the disparity. This trend leads

<sup>16</sup> Maguire, E. A., Gadian, D. G., Johnsrude, I. S., Good, C. D., Ashburner, J., Frackowiak, R. S., & Frith, C. D. (2000). Navigation-related structural change in the hippocampi of taxi drivers. *Proceedings of the National Academy of Sciences*, 97(8), 4398-4403.

<sup>17</sup> Cullen, R. (2001). Addressing the digital divide. *Online Information Review*, 25(5), 311-320.

<sup>18</sup> Haleem, A., Javaid, M., & Singh, R. P. (2022). An era of ChatGPT as a significant futuristic support tool: A study on features, abilities, and challenges. *BenchCouncil transactions on benchmarks, standards and evaluations*, 2(4), 100089. Haleem, A., Javaid, M., & Singh, R. P. (2022). An era of ChatGPT as a significant futuristic support tool: A study on features, abilities, and challenges. *BenchCouncil transactions on benchmarks, standards and evaluations*, 2(4), 100089.

<sup>19</sup> Firat, M. (2023a). How chat GPT can transform autodidactic experiences and open education. Department of Distance Education, Open Education Faculty, Anadolu Unive

to a situation where those already knowledgeable become more advanced, while those with less knowledge fall further behind in competitiveness.

Those with more resources, skills, and positive attitudes toward technology, as well as interest in the content offered, will be more likely to benefit from ChatGPT. Those without these advantages may find the gap between their knowledge and that of others increasing due to the presence of tools like ChatGPT. And A new digital divide is appearing nowadays. According to Common Sense Media, low-income teenagers spend an average of 8 hours and 7 minutes per day using screens for entertainment, while high-income teenagers spend 5 hours and 42 minutes per day. It is paradoxical that Silicon Valley parents are increasingly worried about the impact of screens on their children and are turning to a screen-free lifestyle. Low-income children are addicted to screens, while elite children return to the luxury of wooden toys and human interaction. Just as Chris Anderson said, the digital divide used to be related to access to technology, but now that everyone has access, the new digital divide is reflected in the limitations on acquiring technology. It might be a more urgent thing that improves the quality of using technology in the future.

## 5. Conclusion

Currently, there is a group of people and academics who are calling for a suspension of "training" AI large models, but this move has been met with resistance from others. The former group believes that new technologies bring ethical and professional identity challenges, while the latter emphasizes the need to focus on how to effectively utilize AI technology. Instead of taking sides, it is essential to acknowledge that AI development has become an unstoppable force driving human progress in today's world. And it is also crucial to recognize that the knowledge gap surrounding AI is still prevalent, and underlying issues such as educational inequality and the equitable distribution of social resources persist. No matter how AI develops, it is not about AI itself, it is about our as human how to use AI.

## References

1. Biswas, S. S. (2023a). Role of chat gpt in public health. *Annals of Biomedical Engineering*, 51(5), 868-869.
2. Biswas, S. S. (2023b). Potential use of chat gpt in global warming. *Annals of Biomedical Engineering*, 51(6), 1126-1127.
3. Cullen, R. (2001). Addressing the digital divide. *Online Information Review*, 25(5), 311-320.
4. Firat, M. (2023a). How chat GPT can transform autodidactic experiences and open education. *Department of Distance Education, Open Education Faculty, Anadolu Unive*
5. Firat, M. (2023b). How chat GPT can transform autodidactic experiences and open education. *Department of Distance Education, Open Education Faculty, Anadolu Unive*
6. Gallagher, S. (2013). The socially extended mind. *Cognitive Systems Research*, 25-26, 4-12. <http://doi.org/10.1016/j.cogsys.2013.03.008>
7. Gaziano, C. (2010). Notes on "revisiting the knowledge gap hypothesis: A meta-analysis of thirty-five years of research". *Journalism & Mass Communication Quarterly*, 87(3-4), 615-632.
8. Haleem, A., Javaid, M., & Singh, R. P. (2022). An era of ChatGPT as a significant futuristic support tool: A study on features, abilities, and challenges. *BenchCouncil transactions on benchmarks, standards and evaluations*, 2(4), 100089.
9. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436-444.
10. Maguire, E. A., Gadian, D. G., Johnsrude, I. S., Good, C. D., Ashburner, J., Frackowiak, R. S., & Frith, C. D. (2000). Navigation-related structural change in the hippocampi of taxi drivers. *Proceedings of the National Academy of Sciences*, 97(8),

4398-4403.

11. Persaud, A. (2001). The knowledge gap. *Foreign Affairs*, 107-117.
12. Robson, G. J., & Tsou, J. Y. (2023). *Technology Ethics: A Philosophical Introduction and Readings*. Taylor & Francis.
13. Surameery, N. M. S., & Shakor, M. Y. (2023). Use chat gpt to solve programming bugs. *International Journal of Information Technology & Computer Engineering (IJITC) ISSN: 2455-5290*, 3(01), 17-22.
14. Tichenor, P. J., Donohue, G. A., & Olien, C. N. (1970). Mass media flow and differential growth in knowledge. *Public Opinion Quarterly*, 34(2), 159-170.
15. Weenig, M. W., & Midden, C. J. (1997). Mass-Media Information Campaigns and Knowledge-Gap Effects 1. *Journal of Applied Social Psychology*, 27(11), 945-958.
16. Wei, L., & Hindman, D. B. (2011). Does the digital divide matter more? Comparing the effects of new media and old media use on the education-based knowledge gap. *Mass Communication and Society*, 14(2), 216-235.

## THE NEW AI ERA: AN AMAZING FANTASY(?)

PANAGIOTA RASSIA

*Utrecht University, Utrecht, The Netherlands.  
E-mail: p.rassia@students.uu.nl*

*“And a lean, silent figure slowly fades into the gathering darkness, aware at last that in this world, with great power there must also come — great responsibility!”*

*— Ditko Steve and Stan Lee (1962), Amazing Fantasy #15*

### 1. Introduction

On May 30<sup>th</sup>, the Center for AI Safety released a [single-sentence statement](#), whereby it called attention to the risks posed to humanity by future advancements in AI. Two weeks prior to the release of the warning statement, Sam Altman, the chief executive of OpenAI and one of the statement's signatories, [testified](#) before a Senate panel in Congress. During the hearing, Altman addressed risks posed by generative AI technologies such as ChatGPT and DALL-E, and opined on how these should be tackled. He even suggested that there be a “regulatory intervention by governments [...] to mitigate the risks of increasingly powerful models”.

Concerns regarding actual and potential harms ensuing from the (malicious) deployment of AI technologies have been raised by researchers already in the past.<sup>1</sup> However, it was not up until the launch of GPT-4 in mid-March this year, when proponents and adversaries of AI engaged themselves in animated discussions on whether it is sensible to put a curb on the development of AI. Altman's case is exceptional in that it reifies a paradox: how can someone be genuinely concerned about humanity's fate, yet—at the same time—invest large sums of money in developing and scaling up the size of tools that can have disastrous ramifications for the societal, political and financial stability worldwide? As Matteo Wong recently pointed out in his article “[AI Doomerism Is a Decoy](#)”, the eschatological narrative co-opted by AI companies serves as “a tacit advertisement” for their product—why would one not want to invest in a product powerful enough to induce global-scale changes?—and, moreover, it renders those at the helm of these companies immune from any kind of criticism should things go awry.

Catastrophising on the basis of an illusory sense of intelligence that the performance of generative AI tools creates does not require much effort. It distracts, however, our attention from the actual harms that the (mis)use of such tools is inflicting on individuals and the society at large, such as the perpetuation and reinforcement of discriminatory biases, copyright infringement, as well as environmental, political and financial harms.<sup>2</sup> What is more concerning about the vague scenarios of the imminent robot uprising, is that they often anthropomorphise AI by describing these tools as sentient entities. During the Future Combat Air and Space Capabilities Summit that took place in London this May, the US Air Force's chief of AI Test and Operations, colonel Tucker Hamilton, presented a simulated test in which an AI-assisted drone was trained to destroy a rival air defence system, but, eventually, attacked anyone who attempted

---

<sup>1</sup> Etzioni, A. and Etzioni, O. (2017). Should artificial intelligence be regulated? Issues in Science and Technology (issues.org), Summer; Acemoglu, D. (2021). Harms of AI. Technical report, National Bureau of Economic Research ; Bender, E. M., Gebru, T., McMillan-Major, A., and Smitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? in Proceedings of the 2021 ACM conference on fairness, accountability, and transparency, pages: 610–623.

<sup>2</sup> Acemoglu, D. (2021). Harms of AI. Technical report, National Bureau of Economic Research ; Bender, E. M., Gebru, T., McMillan-Major, A., and Smitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? in Proceedings of the 2021 ACM conference on fairness, accountability, and transparency, pages: 610–623.

to interfere with this command. <sup>3</sup>Commenting on the system's performance, colonel Hamilton stated the following:<sup>4</sup>

- (1) “The system started realising that while they did identify the threat, at times the human
- (2) “AI is also very brittle, i.e. it is easy to trick and/or manipulate.”

Colonel Hamilton's statements exemplify this practice of ascribing human characteristics to tools. In both his statements, colonel Hamilton presents the “system” and “AI” as agents capable of acting on their own in an attempt to absolve those who developed and operated the system from their responsibilities: both the verb realise and the predicates is easy to trick and / or manipulate require that the entity to which they refer be sentient. However, it is not the tools nor a research field that are sentient, but rather the humans who operate and represent them respectively. As Acemoglu<sup>5</sup> points out, AI technologies are not intrinsically malicious; whether these technologies prove to be harmful is highly contingent on the way humans and corporations deploy them. It is crucial, therefore, to distinguish between individuals who act, and tools that are operated / deployed by individuals to assist them in accomplishing their tasks.

This distinction between sentient agents, on the one hand, and assisting tools, on the other, is of particular interest when it comes to assessing the quality of the outputted text produced by AI-assisted text-generating tools such as ChatGPT. More precisely, their ability to produce (seemingly) fluent and coherent text <sup>6</sup>has prompted a series of discussions within the field of Linguistics—as well as other related disciplines—with the main concern being whether these tools do actually understand the text they produce.

In the following section, we elaborate more on this issue by discussing recent empirical evidence by Sinclair and colleagues (2022) suggesting that Large Language Models (LLMs), quite similarly to humans, display structural priming effects, but also dissenting voices claiming otherwise.<sup>7</sup> In section 3, we present initiatives towards educating people on (the appropriate use of) these tools. We argue that, instead of treating technology as a beast we should fear unless it is tamed, people need to learn how to deploy it critically and understand its limitations. As has been repeatedly mentioned throughout this section, AI technologies and AI-assisted tools are not harmful per se. People resort to them on a daily basis and should, therefore, become aware of their full potential: how to reap their benefits, but also avoid possible risks ensuing from their misuse. Section 4 concludes.

## 2. The language conundrum: Is language an ability unique to humans?

The advent of LLMs has challenged a long-standing theory in the field of Linguistics, namely that of the innateness of language. The innateness hypothesis, as this was laid out by Chomsky (1957)<sup>8</sup>, posits that language is an ability unique to humans, and that humans possess a congenital knowledge of linguistic structure. However, the spectacular performance of

---

<sup>3</sup> According to the US Air Force's spokesperson, no such simulation had taken place and “[...] the colonel's comments were taken out of context and were meant to be anecdotal”.

<sup>4</sup> Quoted statements appear as in the source [article](#). Emphasis is added.

<sup>5</sup> Acemoglu, D. (2021). Harms of AI. Technical report, National Bureau of Economic Research, pp.46.

<sup>6</sup> Bender, E. M., Gebru, T., McMillan-Major, A., and Smitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? in *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pp.616.

<sup>7</sup> Bender, E. M., & Koller, A. (2020). Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages: 5185-5198; Bender, E. M., Gebru, T., McMillan-Major, A., and Smitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? in *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages: 610–623.

<sup>8</sup> Chomsky, N. (1957). *Syntactic Structures*. Mouton.



language models such as GPT-4—lauded (e.g., Piantadosi, 2023<sup>9</sup>) and criticised in equal measure—has led researchers to reconsider the validity of this hypothesis.

In a recent study, Sinclair et al.<sup>10</sup> investigated the learning capacities of modern neural language models. More precisely, in order to identify whether language models are capable of learning structural information, they conducted a series of priming experiments. Their experiments were motivated by similar neuro- / psycholinguistic studies revealing that humans are more inclined to produce a sentence of a certain structure if they have been previously presented, i.e., primed, with a sentence of identical structure. Evidence of structural priming lends support to the idea that abstract knowledge about structural information is retained. As far as language models are concerned, this information is crucial for downstream tasks that require Natural Language Understanding (NLU) skills.<sup>11</sup>

The experiments by Sinclair and colleagues examined two syntactic alternations—one in the nominal (dative vs. prepositional phrase alternation) and one in the verbal domain (active vs. passive form alternation)—that allow the same content to be expressed in two (syntactically) distinct ways.<sup>12</sup> Syntactic alternations in the nominal domain included sentences such as A pilot bought an attorney a pie, whose prepositional-phrase alternative is A pilot bought a pie for an attorney; those in the verbal domain involved alternations between the active and passive form of verbs such as The nurse purchased the beer and The beer was purchased by the nurse. In order to verify their hypotheses, Sinclair and colleagues developed the PRIME-LM corpus, a large corpus consisting of approximately 1.3 million prime-target sentence pairs, which they tested on several language models. The results of their study revealed that modern neural language models do exhibit structural priming effects, giving, thus, credence to their hypothesis that models of this type are capable of learning abstract structural information—albeit to a certain extent—as well as recalling this information in order to make informed predictions about the sentences’ structure in follow-up tasks. Nevertheless, as Sinclair and colleagues pointed out, the strength of the priming effect was influenced by the semantic similarity between the prime and target sentence, the proximity with which the prime and target sentences were presented, in addition to the model’s degree of exposure to certain structures during the priming.<sup>13</sup>

The ability of text-generating tools to display human-analogous performance in language production tasks can often beguile us into treating them as ‘thinking’ entities, when, in reality, the ‘success’ of their output relies merely on probabilistic information about the combination of linguistic forms the model has observed in an unfathomable amount of training data.<sup>14</sup> This misconception is further reinforced even by scientific studies on the NLU performance skills of language models, which make imprudent use of the terminology by not indicating explicitly whether these terms are mainly intended as technical terms or they imply that the performance of these models is analogous to human language understanding.<sup>15</sup> As Bender and colleagues (2021) argue, human communication is much more sophisticated in that individuals express communicative intents when interacting with one another.<sup>16</sup> According to Bender and Koller (2020), communicative intents are external to language in the sense that they are grounded in

<sup>9</sup> Piantadosi, S. (2023). Modern language models refute Chomsky’s approach to language. In: *lingbuzz* 007180.

<sup>10</sup> Sinclair, A., Jumelet, J., Zuidema, W., & Fernández, R. (2022). Structural persistence in language models: Priming as a window into abstract language representations. *Transactions of the Association for Computational Linguistics*, 10, pp. 1031-1050.

<sup>11</sup> *ibid* pp.1032

<sup>12</sup> Sinclair, A., Jumelet, J., Zuidema, W., & Fernández, R. (2022). Structural persistence in language models: Priming as a window into abstract language representations. *Transactions of the Association for Computational Linguistics*, 10, pp. 5.

<sup>13</sup> *Ibid* pp. 1039 - 1041

<sup>14</sup> Bender, E. M., Gebru, T., McMillan-Major, A., and Smitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? in *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pp.617.

<sup>15</sup> Bender, E. M., & Koller, A. (2020). Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 5186.

<sup>16</sup> Bender, E. M., Gebru, T., McMillan-Major, A., and Smitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? in *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pp.616.

the real world inhabited by interlocutors.<sup>17</sup> In order, thus, to ascribe human-analogous NLU performance skills to language models, these need not only to possess mastery over the language's structure and its use but also the ability to ground this information in the real world.<sup>18</sup>

### 3. Integrating (generative) AI tools into society and steps forward

As preparation for writing this paper, we decided to participate in a suite of lectures and hands-on workshops centred around the use and applications of GPT-4 that came to our notice. These events are presented in chronological order in Table 1.

Table 1: List with events related to GPT-4 and its applications. The format of the listed events' dates is dd/mm/yyyy..

Event's date	Speaker(s) / Presenter(s)	Event's title	Event's type	Organising institution / association	Target audience
25/05/2023	Dr. Willem Zuidema, University of Amsterdam	The Linguistics of Deep Learning: ChatGPT and Large Language Models	Online lecture	Societas Linguistica Europea (SLE)	University students and researchers of AI or related fields
31/05/2023	Dr. Ana Niño, University of Manchester  Professor Dr. Rudy Loock, University of Lille	Free Online Machine Translators (FOMT) vs. ChatGPT: what opportunities and challenges do these AI tools bring to the language teaching and learning context? Online translators in the classroom: how to empower language learners?	Online lecture	Leiden University Centre for Linguistics (LUCL)	University students and researchers of AI or related fields
06/06/2023	Dr. Crit Cremers	Meaning or what? The semantics of ChatGPT	Hybrid lecture	Leiden University Centre for Digital Humanities (LUCDH)	University students and researchers of AI or related fields
07/06/2023	Dr. Judith Thissen, Utrecht University Dr. Karin van Es, Utrecht University Dr. Rianne van Lambalgen, Utrecht University	ChatGPT – A practical introduction	On-site hands-on workshop	Center for Digital Humanities Utrecht University	University students and researchers of AI or related fields

The initiatives listed above aimed at bringing GPT-4 and its various uses to the broader research community's attention, and fostered a multidisciplinary discussion that is undeniably beneficial. Nevertheless, we should keep in mind that generative AI tools are directed at society at large. Therefore, we suggest that experts should inform and guide the broader public on how to fully and appropriately capitalise on these tools through knowledge events / fairs and seminars that can be hosted in public spaces (e.g., libraries, bookstores, cafés etc.). Finally, in line with

<sup>17</sup> Bender, E. M., & Koller, A. (2020). Climbing towards NLU: On meaning, form, and understanding in the age of data. In Proceedings of the 58th annual meeting of the association for computational linguistics, pp. 5186.


<sup>18</sup> Bender, E. M., & Koller, A. (2020). Climbing towards NLU: On meaning, form, and understanding in the age of data. In Proceedings of the 58th annual meeting of the association for computational linguistics, pp. 5185.

Littman and colleagues<sup>19</sup>, we argue in favour of a bottom-up reform of the educational curriculum, where a core understanding of AI concepts is established in the early stages of education. We hold that not only is this an efficient instructional approach, but also—and most importantly—an ethical and respectful attitude towards future generations, as it will prepare sufficiently AI-literate individuals with a much more elevated sense of responsibility regarding the deployment of AI. These qualities are indispensable in navigating a world in which AI will pervade every aspect of human life.

#### 4. Concluding Remarks

In this paper, we argued that, while regulating AI is indeed necessary in order to ensure that ethical parameters are respected and biases are not perpetuated, this is not a sufficient measure on its own. Many of the existing misconceptions around AI are based on an illusory impression of omnipotence with which AI-assisted systems are commonly associated. However, much like any other tool, AI-assisted tools cannot be detached from their operators. This is also relevant in the case of text-generating tools such as ChatGPT, whose (seemingly) fluent and coherent text production can beguile people into ascribing to these tools human-analogous properties (e.g., reasoning abilities). In order, thus, to ensure that AI-assisted tools are deployed appropriately, humans need not rely only on the convenience these tools provide them, but also (actively seek to) become aware of their limitations and the harms these can inflict. As a final remark, we would like to emphasise that technology should not be regarded in terms of good or bad, nor should it be demonised. People develop and utilise tools to assist themselves in completing seamlessly their daily tasks. Rather than thinking of tools as ‘intelligent’ or ‘powerful’, we should regard them as our ‘superpowers’; ‘superpowers’ that we have to use wisely and, above all, responsibly.

#### References

1. Acemoglu, D. (2021). Harms of AI. Technical report, National Bureau of Economic Research.
2. Bender, E. M., Gebru, T., McMillan-Major, A., and Smitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big?  In Proceedings of the 2021 ACM conference on fairness, accountability, and transparency, pages: 610–623.
3. Bender, E. M., & Koller, A. (2020). Climbing towards NLU: On meaning, form, and understanding in the age of data. In Proceedings of the 58th annual meeting of the association for computational linguistics, pages: 5185-5198.
4. Chomsky, N. (1957). Syntactic Structures. Mouton.
5. Etzioni, A. and Etzioni, O. (2017). Should artificial intelligence be regulated? Issues in Science and Technology (issues. org), Summer
6. Littman, M. L., Ajunwa, I., Berger, G., Boutilier, C., Currie, M., Doshi-Velez, F., Hadfield, G., Horowitz, M. C., Isbell, C., Kitano, H. et al. (2022). Gathering strength, gathering storms: The one hundred year study on artificial intelligence (AI100) 2021 study panel report. arXiv preprint arXiv:2210.15767.
7. Piantadosi, S. (2023). Modern language models refute Chomsky’s approach to language. In: lingbuzz 007180.
8. Sinclair, A., Jumelet, J., Zuidema, W., & Fernández, R. (2022). Structural persistence in language models: Priming as a window into abstract language representations. Transactions of the Association for Computational Linguistics, 10, pages: 1031-1050.

---

<sup>19</sup> Littman, M. L., Ajunwa, I., Berger, G., Boutilier, C., Currie, M., Doshi-Velez, F., Hadfield, G., Horowitz, M. C., Isbell, C., Kitano, H. et al. (2022). Gathering strength, gathering storms: The one hundred year study on artificial intelligence (AI100) 2021 study panel report. arXiv preprint arXiv:2210.15767, pp.71.

## IS GENERATIVE AI DEPRIVING HUMANS OF CREATIVITY?

ALESSIA GIULIMONDI

*Utrecht University, Utrecht, The Netherlands.*

*E-mail: [a.giulimondi@students.uu.nl](mailto:a.giulimondi@students.uu.nl)*

### 1. Introduction

“The operating system of every human culture in history has always been language.” These are the words Harari uses to introduce the explanation of dangers and threats AI technology represents for humanity.<sup>1</sup> According to his view, the fast and uncontrolled deployment of a tool that has gained the ability to efficiently interact with human beings poses an existential threat to human civilization. He explains that “the most important aspect of the current phase of the ongoing AI revolution is that AI is gaining mastery of language at a level that surpasses the average human ability”.<sup>2</sup> This claim seems to agree with Geoffrey Hinton’s recent statements about language models working with backpropagation, an algorithm developed and used for machine learning, initially believed to be a digital replication of how the neural network in the human brain works. Hinton’s belief has now changed and he claims in his recent releases that these language models work very differently and way more efficiently than humans, simply because they rely on more data. “We are very bad at communicating compared with these current computer models that run on digital computers (...) their communication band is huge”.<sup>3</sup> He explains that this is because clones of the same models are able to run on different computers and through these connections they can see “huge amounts of data”. Finally, a recent research on generative AI<sup>4</sup> has shown the escalating potential of persuasion of Large Language Models (LLM) that may have a disruptive impact on society, considering that every individual in both their private and professional activities can use generative AI, just by having a stable Internet connection.

Despite these warnings, very little research focuses on ethical and societal relevance of language use in human cultures and how creating an “alien” intelligence – as Harari called it – that is able to use human language may threaten governments and institutions as we know them. Linguistic knowledge and research is at the core of this crisis, because language models are the tools that are now enabling artificial intelligence to interact effectively with humans. What is most dangerous, according to Harari’s claims, is the ability of AI to gain intimacy with humans, which happens simply through linguistic communication. However, linguistic research has not yet agreed on how language works in communication, how we are able to understand each other and how we even learn to speak in the first place. Nevertheless, algorithms used in machine learning have now released a tool that is able to process human sentences and produce written human-like sentences without neither linguists nor AI programmers knowing exactly how this is happening. As a matter of fact, research in AI and computational linguistics does not have enough information about AI-human interactions and the effects these might have on users.

---

<sup>1</sup> Harari, Y. N. (2018, October). Why Technology Favors Tyranny Artificial intelligence could erase many practical advantages of democracy, and erode the ideals of liberty and equality. It will further concentrate power among a small elite if we don’t take steps to stop it. *The Atlantic*. <https://www.theatlantic.com/magazine/archive/2018/10/yuval-noah-harari-technology-tyranny/568330/>

<sup>2</sup> *ibid*

<sup>3</sup> CBS Mornings. (2023, March 25). Full interview: “Godfather of artificial intelligence” talks impact and potential of AI [Video]. YouTube. <https://www.youtube.com/watch?v=qpoRO378qRY>

<sup>4</sup> Matz, S., Teeny, J., Vaid, S. S., Harari, G. M., & Cerf, M. (2023). The Potential of Generative AI for Personalized Persuasion at Scale.

However, the damages this new advancement in technology can cause might not be avoided through gaining new perspectives and information. The dangers highlighted by Harari, Hinton and well-summarized by Future of Life's open letter are not meant to be solved by more research on specific aspects of AI development.<sup>5</sup> It is worth mentioning Ivan Illich's work, in which he clearly describes how more specialized knowledge and increasing accumulation of scientific research and information in the last decades has led artificial tools invented and built by humans to gain control over them, instead of enabling individuals to make free, better and more efficient use of their own human resources, as it was intended.<sup>6</sup> According to Illich, a systematic analysis of these tools needs to be carried out, in order to understand whether those are depriving humans from their energy and capacity to freely shape their lives. Careful analysis of AI tools and its linguistic interactions with humans are needed to understand the dimension of this threat and whether the benefits actually outweigh the costs.

## 2. An ethical decision

Even though it was briefly described above how language understanding and production represents the turning point of AI training, the ambiguities and risks in which artificial intelligence is being deployed are multifaceted and cover all aspects of society that go far beyond the research expertise of any field. Experts can keep proposing research questions and ideas, pointing out what has not been yet investigated and what deserves closer scrutiny, but this cannot exempt scientists from facing the reality that this decision is ultimately ethical and not technical. The broader research question that I propose to investigate in interdepartmental ethical committees is whether it is truly useful and not harmful in any possible way to human civilization and single individuals financing research in NLP and AI. It is crucial to understand that a pause of six months may not be sufficient to prepare our social and political systems for a technology that the builders themselves are not sure how it works. Moreover, it is of fundamental importance to gain awareness of the threats that such a powerful technology can pose to our democracy if this is built and implemented by private corporations for which the primary purpose of technological development is profit. The potential good that is often used to justify the research of these tools and their implementation does not eliminate the equally dangerous consequences we are already facing now because of these same tools. Scientists should acknowledge the limits that are needed to a safe and healthy technical development and the pace and timing of the discussion on these limits cannot be dictated by a capitalist agenda of private corporations.<sup>7</sup>

## 3. Language deprivation and human creativity

It is often said that what distinguishes humans from machines and AI is creativity. And creativity is always achieved by means of language, shaping our understanding of both internal and external world, defining the borders of reality through communication and thinking, thinking and consequent action.<sup>8</sup> When language is missing, cognitive skills are not fully developed<sup>9</sup> and poetic realization (i.e. creativity) is likely to be reduced as well.

---

<sup>5</sup> Future of Life Institute. (2023, May 5). Pause Giant AI Experiments: An Open Letter - Future of Life Institute. <https://futureoflife.org/open-letter/pause-giant-ai-experiments/>

<sup>6</sup> Illich, I. (1973). Tools for Conviviality. Perennial Library.

<sup>7</sup> CBS Mornings. (2023, March 25). Full interview: "Godfather of artificial intelligence" talks impact and potential of AI [Video]. YouTube. <https://www.youtube.com/watch?v=qpoRO378qRY>.

<sup>8</sup> Kozulin, A. (2018). Mediation and internalization. Conceptual analysis and practical applications. In J.P. Lantolf, M. E. Poehner & M. Swain (Eds.), *The Routledge handbook of sociocultural theory and second language development* (pp. 23-41). Routledge; Lantolf, J. P., Poehner, M. E., & Thorne, S. L. (2020). Sociocultural theory and L2 development. In *Theories in second language acquisition* (pp. 223-247). Routledge.

<sup>9</sup> Cheng, Q., Halgren, E., & Mayberry, R. (2018). Effects of early language deprivation: Mapping between brain and behavioral outcomes. In *Proceedings of the 42nd annual Boston University conference on language development* (pp. 140-152). Somerville, MA: Cascadia Press.

However, in chess competitions, creativity is considered a hallmark of AI behavior,<sup>10</sup> a behavior independently developed from human training, but only founded on human data. The endless collection of human information today stored in a few large databases represents a substantial threat to human civilization. LLM are based on the presupposition that human high mental functions (e.g. philosophical thinking), which according to Vygotsky's theory happen through the mediation of language,<sup>11</sup> can be modeled, thus, predicted and finally swayed. Persuasion has always happened in human history through the shaping of language. Now persuasion is carried out by an alien intelligence of which mechanisms we do not know, upon which we cannot ultimately have any control. While humans are losing their own language in a game of pantomime and persuasion, an unknown intelligence is acquiring deep understanding of that same language, gaining human trust and, therefore, an escalating power of influencing and leading human thoughts, ideas and decisions.<sup>12</sup>

This was all made possible by a persistent conviction characterizing the past two centuries that humans are fallible and technical development is the only solution to humans' proclivity to failure. Contemporary corruption of language, well-described by Illich (1973), shows this belief. Standardized and pre-fabricated language is nowadays the common practice of language use. This implicitly forces individuals to impoverish the complexity of their thoughts and trivialize the poetic value of their intuitions. This is because the more human language resembles machine's behavior, the more human mind is easy to control. The more obvious and transparent human thinking becomes, the more docile and easy to govern it is. Indeed, humans ceded creativity to machines, in turn of a more predictable behavior. Humans devoid of their poetic capability, made possible through language mediation, are not different from any other tool that can be exploited for industrial purposes. Illich identified this process as the industrialization of man, made possible through the industrialization of language. Nominal language, characteristic of scientific language, is language deprived from its creativity. Purely descriptive, constantly striving for transparency and objectivity, nominal language is serving industrial purposes and, by extension, purposes of machines, which are far from serving human needs. Given this overview, it is not surprising that last decades were marked by a growing feeling of competition with machines. Indeed, Harari opens his article describing an ominously familiar sense of irrelevance most people in industrialized countries experience every day. It is increasingly clear to members of our society how irrelevant human actions are if compared to the efficiency of automated machines. According to the mainstream views of human history, human development is in its essence a series of technical developments. Therefore, AI development comes unquestioned and it is easy to be considered inevitable and unstoppable. The necessity of technical development has never been seriously called into doubt, since this is considered the core of our civilization and the only reasonable solution to the "human problem". Indeed, humans are emotional, forgetful and egocentric. Machines are now fixing these human bugs.

According to Vaid, the more humans are made irrelevant in the development of society, the more societal problems will be solved. The more humans acknowledge their own irrelevance, the safest the planet will be. Depriving humanity from its sense of dignity through, first, language deprivation and, then lending that same language to machines is the most well-succeeded attempt to put humans to one side of their own history and deliver the leadership to a non-human intelligence.

Indeed, what is left to humans if the fundamental tool for creativity, which is language, has lost its power to define humanity and it is now fed to a machine that outperforms its creator? If language cannot be recovered as a fundamental tool to fight the pervasive sense of

---

<sup>10</sup> Harari, Y. N. (2018, October). Why Technology Favors Tyranny Artificial intelligence could erase many practical advantages of democracy, and erode the ideals of liberty and equality. It will further concentrate power among a small elite if we don't take steps to stop it. *The Atlantic*.

<sup>11</sup> Kozulin, A. (2018). Mediation and internalization. Conceptual analysis and practical applications. In J.P. Lantolf, M. E. Poehner & M. Swain (Eds.), *The Routledge handbook of sociocultural theory and second language development* (pp. 23-41). Routledge.

<sup>12</sup> Matz, S., Teeny, J., Vaid, S. S., Harari, G. M., & Cerf, M. (2023). *The Potential of Generative AI for Personalized Persuasion at Scale*.

meaninglessness of our times and growing perception of irrelevance already accepted by young generation, we are now already facing the end of human civilization.

Indeed, young generations are used to live in this sense of apocalypse. They are not shocked by their own irrelevance. A lukewarm concern and foggy lines of reasoning is what is left to my generation. Intuitions deprived of their means of expression are the only hopes few ambitious young people can cling to in order not to drown in this common acknowledgment of human marginality our society has embraced.<sup>13</sup>

Seeds of poetic realization have to be recovered through language, through a radical refuse of its modeling. Modeling presupposes the predictability of language, which is, in this way, restricted to mere assembling of constituents, through specific mechanisms that can be mimicked and improved by machines which have more energetic resources than humans.

For this reason, the decisions we need to make are not bound to scientific discoveries, nor to more research and more experiments.<sup>14</sup> The decisions are ethical and are not to be delegated to some experts' knowledge. The threat is posed to human civilization and the community of humans as a whole have to tackle it.

## References

1. CBS Mornings. (2023, March 25). *Full interview: "Godfather of artificial intelligence" talks impact and potential of AI* [Video]. YouTube. <https://www.youtube.com/watch?v=qpoRO378qRY>
2. Cheng, Q., Halgren, E., & Mayberry, R. (2018). Effects of early language deprivation: Mapping between brain and behavioral outcomes. In *Proceedings of the 42nd annual Boston University conference on language development* (pp. 140-152). Somerville, MA: Cascadia Press.
3. Future of Life Institute. (2023, May 5). *Pause Giant AI Experiments: An Open Letter - Future of Life Institute*. <https://futureoflife.org/open-letter/pause-giant-ai-experiments/>
4. Illich, I. (1973). *Tools for Conviviality*. Perennial Library.
5. Kozulin, A. (2018). Mediation and internalization. Conceptual analysis and practical applications. In J.P. Lantolf, M. E. Poehner & M. Swain (Eds.), *The Routledge handbook of sociocultural theory and second language development* (pp. 23-41). Routledge.
6. Lantolf, J. P., Poehner, M. E., & Thorne, S. L. (2020). Sociocultural theory and L2 development. In *Theories in second language acquisition* (pp. 223-247). Routledge.
7. Matz, S., Teeny, J., Vaid, S. S., Harari, G. M., & Cerf, M. (2023). The Potential of Generative AI for Personalized Persuasion at Scale.
8. Harari, Y. N. (2018, October). Why Technology Favors Tyranny Artificial intelligence could erase many practical advantages of democracy, and erode the ideals of liberty and equality. It will further concentrate power among a small elite if we don't take steps to stop it. *The Atlantic*. <https://www.theatlantic.com/magazine/archive/2018/10/yuval-noah-harari-technology-tyranny/568330/>
9. Yuval Noah Harari. (2023, May 14). *AI and the future of humanity | Yuval Noah Harari at the Frontiers Forum* [Video]. YouTube. <https://www.youtube.com/watch?v=LWiM-LuRe6w>

---

<sup>13</sup> Harari, Y. N. (2018, October). Why Technology Favors Tyranny Artificial intelligence could erase many practical advantages of democracy, and erode the ideals of liberty and equality. It will further concentrate power among a small elite if we don't take steps to stop it. *The Atlantic*. <https://www.theatlantic.com/magazine/archive/2018/10/yuval-noah-harari-technology-tyranny/568330/>

<sup>14</sup> Illich, I. (1973). *Tools for Conviviality*. Perennial Library.





## AUTHOR INDEX

- Arnold, J. 64
- Ballester, L. S. 109
- Breemen, K. 123
- Breemen, V. 123
- Bringsjord, A. 5
- Bringsjord, S. 5
- Broos, L. 126
- Calboli, S. 55
- Callari, T. C. 79
- Chippendale, P. 109
- Couceiro, M. 109
- de Cock Buning, M. 3
- de Jong, E. 124
- Dimitropoulos, N. 90
- Eimontaite, I. 64, 90, 99
- Ferrari, F. 125
- Ferreira, M. I. A. 47
- Fletcher, S. R. 90, 99, 109
- Giancola, M. 28
- Giulimondi, A. 152
- Godhania, S. 99
- Govindarajulu, N. S. 5
- Hubbard, E-M. 79
- Israeli, I. 90
- Kleijssen, J. 4
- Liu, J. 141
- Lohse, N. 79
- Makris, S. 90
- Mawle, R. 16
- McGloin, H. 16
- Michalos, G. 90
- Mongelli, M. 39
- Ogundele, P. 64
- Oostveen, A-M. 109
- Operto, F. 39
- Ospina, A. P. C. 99
- Oswald, J. 28
- Rassia, P. 147
- Ruijter, E. 13
- Ruiz, E. 99
- Sanders, M. 13
- Sashidharan, V. 90
- Scott-Parker, S. 6
- Segura, A. G. 99
- Steer, Z. 64
- Studley, M. 9, 16
- Świerzawska, L. 129
- Tucker, S. 90
- van Dijck, J. 13
- Veruggio, G. 39
- Wachter, T. 135
- Winfield, A. 16





