Designing Positive AI

How optimizing for contextual wellbeing inspired a design method for artificial intelligence that promotes human flourishing

van der Maden, W.L.A.

DOI
10.4233/uuid:7a341d93-3a51-4df2-9d0a-fcade9008e63

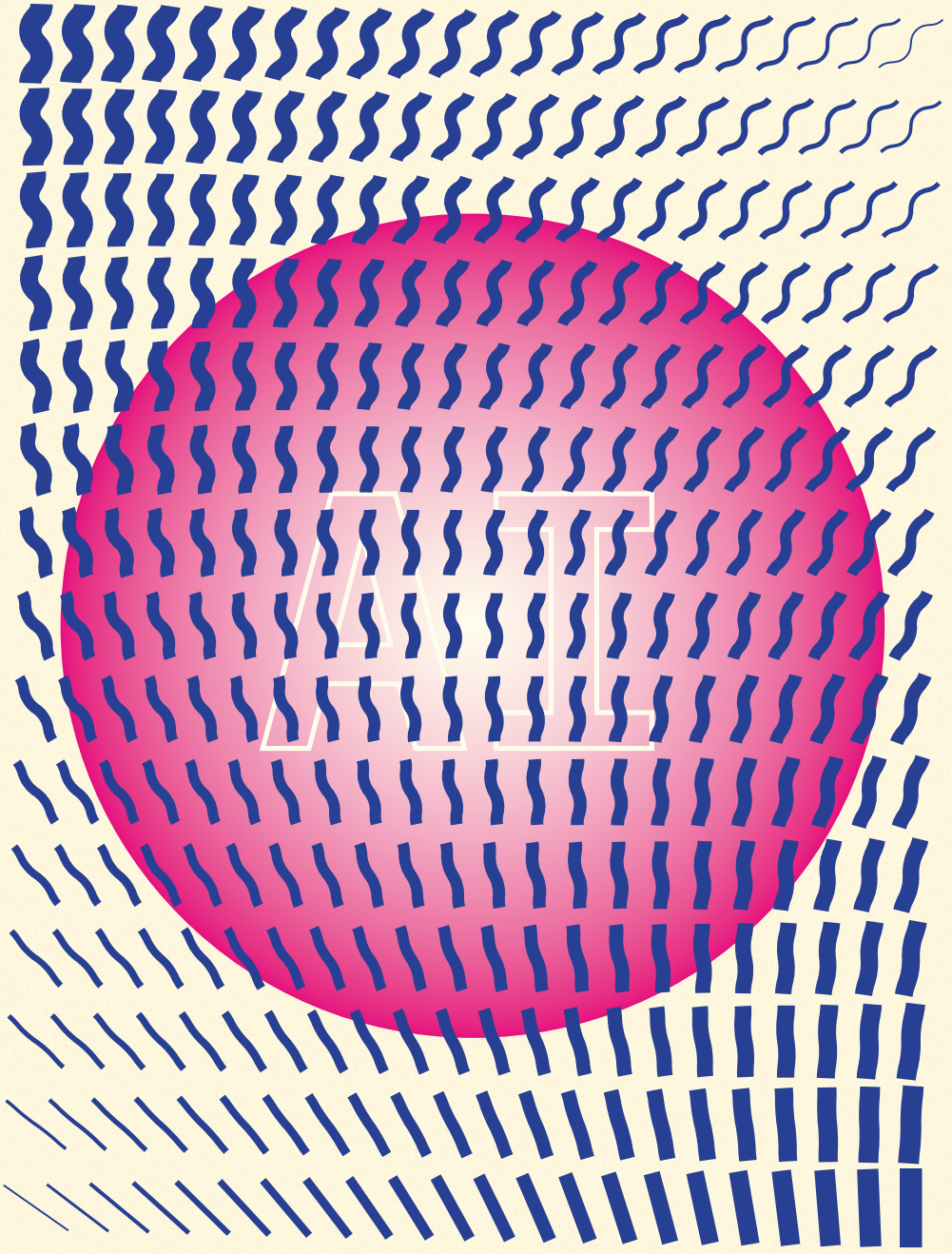
Publication date
2024

Document Version
Final published version

Citation (APA)
van der Maden, W. L. A. (2024). *Designing Positive AI: How optimizing for contextual wellbeing inspired a design method for artificial intelligence that promotes human flourishing*. [Dissertation (TU Delft), Delft University of Technology]. https://doi.org/10.4233/uuid:7a341d93-3a51-4df2-9d0a-fcade9008e63

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Designing Positive AI



Willem van der Maden

How optimizing for contextual wellbeing inspired a design method for artificial intelligence that promotes human flourishing

Designing *Positive AI*

How optimizing for contextual wellbeing inspired a design method for artificial intelligence that promotes human flourishing

Dissertation

for the purpose of obtaining the degree of doctor
at Delft University of Technology
by the authority of the Rector Magnificus prof. dr. ir. T.H.J.J. van der Hagen;

Chair of the Board for Doctorates
to be defended publicly on Wednesday 8 May 2024 at 15:00 o'clock

by

Willem Lennert Antoon VAN DER MADEN

Master in Psychology,
Utrecht University, Nederland,
born in Boxmeer, The Netherlands.

This dissertation has been approved by the promotors.

Composition of the doctoral committee:

Rector Magnificus,	chairperson
Prof. dr. P.P.M. Hekkert	Delft University of Technology, *promotor*
Dr. J.D. Lomas	Delft University of Technology, *copromotor*

Independent members:

Prof. dr. J. Forlizzi	Carnegie Mellon University
Prof. dr. R. Calvo	Imperial College London
Prof. dr. L. Chen	Eindhoven University of Technology
Prof. dr. ir. I.R. van de Poel	Delft University of Technology
Prof. dr. ir A. Bozzon	Delft University of Technology

Keywords: design, wellbeing, artificial intelligence, cybernetics, positive psychology

Printed by: Ipskamp Printing, Enschede

Front & Back: Chanelle Hool

ISBN 978-94-6473-488-1

An electronic version of this dissertation is available at
http://repository.tudelft.nl/.

Contents

Contents v

Summary ix

Samenvatting xv

1 Introduction 1
- 1.1 Current research: Positive AI . . . 4
- 1.2 Research aim . . . 4
- 1.3 Research questions . . . 5
- 1.4 Research approach . . . 6
- 1.5 How to read this dissertation . . . 7

2 Background 9
- 2.1 Artificial Intelligence . . . 10
  - 2.1.1 Cybernetics . . . 11
  - 2.1.2 Ethical AI . . . 15
  - 2.1.3 AI Alignment . . . 17
- 2.2 Wellbeing . . . 19
  - 2.2.1 Hedonia & Eduaimonia . . . 20
  - 2.2.2 Wellbeing as happiness . . . 21
  - 2.2.3 Wellbeing as flourishing & Positive Psychology . . . 23
  - 2.2.4 'Second-wave' Positive Psychology . . . 23
  - 2.2.5 'Third-wave' Positive Psychology . . . 26
  - 2.2.6 Can wellbeing be experienced by communities? . . . 27
  - 2.2.7 Assessing wellbeing . . . 29
  - 2.2.8 Why should wellbeing be an objective for AI? . . . 31
  - 2.2.9 Wellbeing metrics in AI . . . 32
- 2.3 Intermezzo: How might AI affect wellbeing? . . . 34
- 2.4 The role of human-centered design . . . 36
  - 2.4.1 Positive Design, Technology & Computing . . . 38

3 Seven key challenges for designing Positive AI 41
3.1 Introduction . . . 42
3.1.1 Ethical AI . . . 43
3.1.2 AI Alignment . . . 43
3.1.3 Why human-centered design? . . . 44
3.1.4 Why wellbeing? . . . 45
3.1.5 Framing the challenges . . . 46
3.2 Key Challenges . . . 48
3.2.1 Challenges related to modeling wellbeing . . . 51
3.2.2 Challenges related to assessing wellbeing . . . 55
3.2.3 Challenge related to designing for wellbeing . . . 58
3.2.4 Challenges related to optimizing wellbeing . . . 60
3.3 Discussion . . . 62
3.3.1 Limitations and final remarks . . . 66

4 Case study: My Wellness Check 67
4.1 Introduction . . . 68
4.2 Related work: designing AI for wellbeing . . . 71
4.2.1 Cybernetics: a conceptual framework . . . 72
4.3 Case study: My Wellness Check . . . 76
4.3.1 Theoretical approach to wellbeing . . . 77
4.3.2 Community-led survey design . . . 78
4.3.3 Deploying the assessment of community wellbeing . . . 80
4.4 Qualitative results . . . 83
4.5 Designing for community action . . . 83
4.5.1 Experimental evaluation . . . 88
4.5.2 Experimental results . . . 90
4.6 Discussion . . . 91
4.6.1 Vision: designing AI for community wellbeing . . . 92
4.6.2 Limitations . . . 94
4.6.3 Design thinking, AI thinking, and cybernetic thinking 95
4.7 Conclusions . . . 97

5 Developing and evaluating a method for Positive AI 99
5.1 Introduction . . . 100
5.2 Background . . . 103
5.2.1 Artificial Intelligence . . . 103
5.2.2 Cybernetics: AI as sociotechnical system . . . 104
5.2.3 Challenges of *designing AI* . . . 105
5.2.4 Challenges of designing AI *for Wellbeing* . . . 106

5.3 A Design Method for Positive AI . . . 109
5.3.1 Development of the method . . . 109
5.3.2 Phases of the Positive AI Method . . . 110
5.3.3 Method applied to fictional example of a streaming platform . . . 118
5.4 Multiple-case study . . . 120
5.4.1 Procedure . . . 121
5.4.2 Materials: Design outcomes . . . 122
5.4.3 Results of the case studies . . . 126
5.5 Narrative-based study with experts . . . 129
5.5.1 Method . . . 130
5.5.2 Results of the narrative-based study . . . 131
5.6 Discussion . . . 131
5.6.1 Reflections on the case-study: comparing efficacy and usability . . . 133
5.6.2 Reflections on the expert study: assessing impact, desirability, and feasibility . . . 134
5.6.3 Limitations of the study . . . 135
5.6.4 The method & existing approaches . . . 136
5.6.5 Proposed adaptations & opportunities for future work 138
5.7 Conclusion . . . 139

6 General discussion & Conclusion 141
6.1 Research questions . . . 142
6.2 Summary and contributions to research and design . . . 142
6.3 Recommendations . . . 144
6.3.1 Integrate HCD in AI development cycles . . . 145
6.3.2 Balance short-term needs with long-term wellbeing . . 149
6.3.3 Model and measure wellbeing in context . . . 152
6.3.4 Establish multiple feedback loops . . . 157
6.3.5 Focus on flourishing not merely mitigating harm . . . 162
6.3.6 Positive AI is a moving target, not an endpoint . . . 163
6.4 Further considerations . . . 166
6.4.1 Connecting the method to the broader field of AI . . . 166
6.4.2 Reflecting on wellbeing as orienting metric . . . 168
6.4.3 Limitations . . . 170
6.5 Conclusion . . . 173

Bibliography 175

Appendix **213**

Acknowledgements **225**

Curriculum Vitæ **227**

List of Publications **229**

Summary

AI's rise, from *curatorial* AI in *YouTube* to *generative* AI in *ChatGPT*, shows potential for both progress and harm. Adopting a Positive Design approach aimed at directly enhancing human wellbeing, this dissertation develops the concept of *Positive AI*. Key questions address how wellbeing manifests in AI systems, how it can be measured, how to design interventions, and how to evaluate them. Outcomes include conceptual frameworks, case studies demonstrating approaches, proposed methods, and evaluations. This culminates in recommendations to further mature these nascent perspectives and capabilities towards AI that actively cultivates human flourishing.

Theoretical Framework and Challenges

This dissertation introduces a cybernetics perspective (Section 2.1.1) to organize the challenges for designing AI for wellbeing (Chapter 3). Within this perspective, 'sensors' refer to components adept at measuring wellbeing indicators, while 'actuators' denote elements that respond to these measurements to improve outcomes. Connecting sensors and actuators creates assessment-action feedback loops that promote continual alignment of systems with community wellbeing goals. Underscoring the constructivist nature of design, cybernetics foregrounds communication, ethical responsibility, and the social construction of meaning. It provides theoretical grounding for human-centered design that connects ethics and interaction through an emphasis on circular causality and the designer's role as an active observer. This supports designing systems capable of nuanced, contextualized assessments and interventions to cultivate wellbeing.

Consequently, seven key challenges (1-7) can be organized around four key questions:

1. **Modeling the state of the system:** How do we operationally define wellbeing within the context of a particular sociotechnical system?
    - For example: What wellbeing dimensions are important (1) in the context of *Netflix* and how do we attribute changes (2) in wellbeing to components of the system?

2. **Assessing the state of the system:** How do we translate qualitative experiences into assessment metrics?

    - For example: How can we elicit (3) how people feel about their interactions with *TikTok*, and how can these experiences be translated into metrics (4) that can be used for assessing future interventions and optimization processes?

3. **Designing system actuators:** How do we design interventions in AI systems that promote and enhance wellbeing?

    - For example: How do we know where in the sociotechnical system of *ChatGPT* we should and can intervene (5), and how do we know whether our potential interventions will achieve the desired effect?

4. **Optimizing the system objective:** How do we know whether we are getting close to our desired goal?

    - For example: How might we manage tradeoffs (6) between autonomy and social connection in designing for wellbeing on *Reddit*, and how do align immediate outcomes with long-term wellbeing goals (7)?

Longitudinal Case-study

Providing practical demonstrations of these challenges in context, Chapter 4 describes a case study at Delft University of Technology, presenting "My Wellness Check," a cybernetic system for community wellbeing during COVID-19. The project, spanning two years, engaged 20,311 participants in seven iterative studies, each designed to enhance the assessment tool's relevance to specific wellbeing needs. Notably, one of these studies, involving 1,719 participants, focused on a comparison experiment. This experiment contrasted a globally validated wellbeing instrument, a domain-specific instrument, and the project's contextualized instrument, with the latter showing favorable outcomes. The design process of both the system "sensors" and its "actuators" was highly participatory, integrating feedback from students, staff, and other stakeholders. This community-led approach, alongside context-sensitive wellbeing assessments, formed a feedback loop that guided organizational actions and enhanced wellbeing, demonstrating the application of cybernetic principles in a complex sociotechnical context.

Developing and Evaluating a Design Method for Positive AI

Seeking to address the key challenges outlined above, and informed by the empirical insights gained from the case study, Chapter 5 presents a novel approach to integrate wellbeing into AI design. This method, motivated by the gap in translating wellbeing concepts into AI design, aims to foster AI systems that respect and enhance human values and wellbeing. It employs a cybernetic approach within an iterative, collaborative framework, involving stakeholder feedback for continual refinement. The method systematically integrates wellbeing into AI design through distinct phases, each focusing on different aspects from contextualizing wellbeing needs to the continuous alignment of AI behavior with wellbeing goals.

1. **Contextualize**: Understanding wellbeing in specific contexts, considering the complex, multi-faceted nature of wellbeing and how it manifests in different settings.

2. **Operationalize**: Transforming the abstract model of wellbeing into actionable, measurable metrics. This phase involves refining the understanding of wellbeing and making the concepts tangible for application in AI design.

3. **Design & Prototype**: Developing AI interactions that aim to enhance wellbeing. This phase involves utilizing insights from the previous stages to ideate and create design strategies that align with wellbeing goals.

4. **Test & Implement**: Implementing the optimized interactions conceptualized in the design phase. This phase requires a collaborative effort from designers and engineers to realize the envisioned designs.

5. **Restart**: A continuous alignment process that revisits the contextual understanding and wellbeing model, ensuring ongoing relevance and alignment with evolving user needs and technological capabilities.

The effectiveness of the method was exemplified through three distinct student projects, each applying the approach to a different domain: dating, nutrition, and music streaming. These projects effectively demonstrated the method's versatility in guiding AI development with a focus on wellbeing. Key insights and learning outcomes from these projects were synthesized to refine and improve the methodology. Expert evaluations of the projects, considering design quality, technical feasibility, and wellbeing enhancement

potential, provided substantial evidence of the method's practicality and efficacy in real-world applications.

Recommendations

Finally, chapter 5 synthesizes insights from across the research into a set of recommendations and reflections to advance the Positive AI agenda. The recommendations can be summarized as follows:

1. **Integrate HCD methods into AI cycles.** Wellbeing's experiential nature means HCD principles like understanding human experiences, taking a systemic perspective, and iterative improvement can align AI with real-world impacts on users. HCD reveals diverse needs, enables ecosystem interventions, and supports gradual adaptation through prototyping. Overall, focusing design on human experiences makes HCD essential for developing AI that fosters wellbeing.

2. **Balance immediate desires and long-term wellbeing in AI.** Optimization often meets transient needs but struggles to support enduring fulfillment. AI should develop mechanisms to understand timeframes' interplay on human flourishing. Contextual metrics alongside wellbeing assessments reveal optimal tradeoff scenarios. Rather than an "AI nanny," research on aligning gratification and fundamental needs could enable wiser applications. Examples show design can resolve dilemmas by naturally aligning behaviors with interests. Overall, AI should consider transient impacts and lifelong trajectories in promoting wellbeing.

3. **Model wellbeing in context by blending theories with local insights.** Contextual modeling of wellbeing, integrating universal theories with local insights, not only tailors interventions to community-specific needs, enhancing effectiveness and relevance but also ensures adaptability to cultural and technological changes, maintaining intervention relevance over time. This approach enables the precise attribution of wellbeing changes to interventions, fostering dynamic support tailored to immediate needs. Additionally, it facilitates a nuanced translation between quantitative data and qualitative experiences, grounding interventions in the rich, lived realities of individuals. Such a comprehensive understanding of wellbeing supports the development of interventions that are deeply resonant and impactful, fostering genuine human flourishing across diverse contexts.

4. **Establish multi-layer feedback loops.** Weaving together the various facets of sociotechnical systems requires the establishment of multi-layer feedback loops. For example, by coupling qualitative insights with system metrics, there is a mechanism in place to scale up the conversation and make the translation between lived experiences and the system as a whole. By establishing many such loops—the recommendation discusses, for instance, the benchmarking loop, product development loop, and business loop—we ensure that the relevant facets of the system are included in the decision-making process. Engaging these feedback channels throughout the design process enables reciprocal learning between systems, users, designers, and other relevant stakeholders, fostering a continuous cycle of improvement and alignment.

5. **Shift from mitigating harm to actively cultivating human flourishing.** AI design often just avoids adverse impacts, but harm reduction alone cannot guarantee wellbeing. Positive AI applies Positive Design's strength-based philosophy of enhancing lives rather than solely removing negatives. It proactively seeks opportunities like educational features tailored to interests. However, unintended consequences show the need for iteration. While complementary, Positive AI's forward-looking approach more fully realizes potential versus reactive stances. Ultimately, designing systems to actively support thriving creates positive change beyond what harm avoidance alone achieves.

6. **Positive AI is an ongoing process, not an endpoint.** Wellbeing's fluidity, inevitable alignment tradeoffs, and continuously evolving sociotechnical systems mean developers must persistently re-evaluate impacts and adapt systems to dynamic user needs. Rather than superficial "ethics washing," achieving positive influence requires authentic commitment to iterative realignment. As contexts shift, so must systems remain supportive. Overall, continual assessments and modifications enable incremental progress versus treating alignment as a static box to check.

The final section explores connections between the proposed Positive AI method and the broader field of AI, discussing future steps like empirical validation through industry collaborations and expanding diverse case studies across contexts. It also reflects on using wellbeing as the orienting metric, including merits like accounting for values contributing to human

flourishing but also limitations like oversimplifying value complexity or ignoring non-human entities. Additionally, it acknowledges key limitations of the research itself, such as the lack of industry implementations and guidance on equitable participation. The section also examines the need to further explore unintended consequences and proactive ethical mitigation strategies. Overall, through multifaceted efforts spanning advocacy, policy, and community building, the Positive AI agenda can progressively guide innovation trajectories towards enhancing societal wellbeing.

Conclusion

In closing, this dissertation makes important headway in charting a path towards AI that actively prioritizes human wellbeing. While further efforts are needed to fully realize Positive AI's potential, the concepts, methods, and recommendations presented aim to spur reflection and progress in steering innovation trajectories toward supporting societal flourishing.

Samenvatting

De opkomst van Kunstmatige Intelligentie—hierna “Artificial Intelligence (AI)”—brengt zowel aanzienlijke mogelijkheden als risico’s met zich mee. Daarom is het cruciaal om juist *nu* AI-ontwikkelingen te sturen richting blijvend maatschappelijk voordeel. Vanuit de benadering van *Positive Design,* gericht op het direct verbeteren van het menselijk welzijn, ontwikkelt dit proefschrift het concept van *Positive AI.* Het verkent de rol van ontwerpers bij het sturen van AI-innovaties richting het holistisch ondersteunen van menselijke welbevinden, en niet alleen het optimaliseren van winst of gebruikersbetrokkenheid. Door middel van mensgerichte methoden streeft dit onderzoek naar het vergroten van kennis over en het verbeteren van technieken voor het meten van de impact van AI op welzijn, om zo interventies te ontwerpen die deze impact iteratief verbeteren. Centraal hierbij zijn vragen die gaan over hoe welzijn zich manifesteert in AI-systemen, hoe dit gemeten kan worden, hoe positieve interventies ontworpen kunnen worden, en hoe deze te evalueren zijn. De uitkomsten van het onderzoek, waaronder conceptuele kaders, casestudies, en ontwerpmethoden, leggen samen een solide basis voor AI die menselijke bloei bevordert. Dit leidt tot een reeks aanbevelingen bedoeld voor onderzoekers en professionals in verschillende vakgebieden. Deze ontluikende inzichten en vaardigheden kunnen zo verder worden ontwikkeld in de richting van AI die een blijvend positieve impact heeft op de wereld.

Theoretisch Kader en Uitdagingen

Dit proefschrift begint met de introductie van een cybernetisch kader (Sectie 2.1.1) om de uitdagingen voor het ontwerpen van AI voor welzijn te organiseren (Hoofdstuk 3). Binnen deze context verwijzen ‘sensoren’ naar componenten die bedreven zijn in het meten van welzijnsindicatoren. ‘Actuatoren’ daarentegen, duiden elementen aan die reageren op deze metingen om de resultaten te verbeteren. Het koppelen van sensoren en actuatoren creëert feedbacklussen die een continu proces van evaluatie en actie mogelijk maken en zo de afstemming van systemen op gemeenschappelijke welzijnsdoelen bevorderen. Door de nadruk te leggen op de constructivistische aard van ontwerp, brengt cybernetica communicatie, ethische verantwoordelijkheid

en de sociale constructie van betekenis naar voren. Het biedt theoretische onderbouwing voor mensgericht ontwerp dat ethiek en interactie verbindt door de nadruk te leggen op circulaire causaliteit en de rol van de ontwerper als actieve waarnemer. Binnen de context van Positive AI, helpt cybernetica bij het aanpakken van een aantal uitdagingen waarmee hedendaagse AI wordt geconfronteerd door het integreren van interdisciplinaire perspectieven die de voorrang geven aan menselijke betekenisgeving. Dit ondersteunt het ontwerpen van systemen die in staat zijn tot genuanceerde welzijnsevaluaties en AI-interventies om welzijn te cultiveren.

Hierop volgend kunnen de zeven belangrijkste uitdagingen (1-7) rond vier kernvragen worden gecategoriseerd:

1. Het **conceptualiseren** van de staat van het systeem: Hoe definiëren we welzijn binnen de context van een bepaald sociotechnisch systeem?
    - Bijvoorbeeld: Welke dimensies van welzijn zijn belangrijk (1) in de context van *Netflix* en hoe wijzen we veranderingen (2) in welzijn toe aan componenten van het systeem?

2. Het **beoordelen** van de staat van het systeem: Hoe vertalen we kwalitatieve ervaringen naar beoordelingsmetrieken?
    - Bijvoorbeeld: Hoe kunnen we achterhalen (3) hoe mensen zich voelen over hun interacties met *TikTok*, en hoe kunnen deze ervaringen worden vertaald naar metrieken (4) die gebruikt kunnen worden voor het beoordelen van toekomstige interventies en optimalisatieprocessen?

3. Het **ontwerpen** van systeemactuatoren: Hoe ontwerpen we interventies in AI-systemen die welzijn bevorderen en versterken?
    - Bijvoorbeeld: Hoe weten we waar in het sociotechnische systeem van *ChatGPT* we moeten en kunnen ingrijpen (5), en hoe weten we of onze potentiële interventies het gewenste effect zullen bereiken?

4. Het **optimaliseren** van het systeemdoel: Hoe weten we of we dichter bij ons gewenste doel komen?
    - Bijvoorbeeld: Hoe kunnen we omgaan met afwegingen (6) tussen autonomie en sociale verbinding bij het ontwerpen voor welzijn op *Reddit*? En hoe kunnen we onmiddellijke resultaten afstemmen op langetermijndoelstellingen (7) voor welzijn?

Longitudinale Casestudy

Om deze uitdagingen in de praktijk te demonstreren, beschrijft Hoofdstuk 4 een casestudy aan de Technische Universiteit Delft, waarbij "My Wellness Check" wordt gepresenteerd, een cybernetisch systeem voor gemeenschapswelzijn tijdens COVID-19. Het project duurde twee jaar betrok en 20.311 deelnemers in zeven iteratieve studies elk ontworpen om de relevantie van het beoordelingsinstrument voor specifieke welzijnsbehoeften te verbeteren. Een van deze studies was bijvoorbeeld een vergelijkend experiment met 1.719 participanten. Dit experiment vergeleek een wereldwijd gevalideerd welzijnsinstrument, een domeinspecifiek instrument en het gecontextualiseerde instrument van het project. Uit de vergelijking bleek dat het gecontextualiseerde instrument van het project, de meest gunstige resultaten toonde—wat insinueerde dat de contextuale aanpak succesvol was. Het ontwerpproces van zowel de 'sensoren' van het systeem als de 'actuatoren' was zeer participatief en integreerde feedback van studenten, personeel en andere belanghebbenden. Dit initieerde een feedbacklus voortkomend uit onze gemeenschapsgerichte benadering verrijkt met contextgevoelige welzijnsbeoordelingen. Deze lus, bestaande uit gerichte acties voortkomend uit systeemwijde welzijnsevaluaties, stuurde zo de institutionele acties van de TU Delft richting het gemeenschappelijke doen van welzijn. Dit proces kan worden gezien als een praktische demonstratie van de toepassing van cybernetische principes binnen een complex sociotechnisch systeem.

Ontwikkelen en Evalueren van een Ontwerpmethode voor Positive AI

Geïnformeerd door de empirische inzichten uit de casestudy, introduceert Hoofdstuk 5 een nieuwe methode om welzijn te integreren in het ontwerp van AI als antwoord op de eerdergenoemde uitdagingen. Het benadrukt de noodzaak om AI af te stemmen op complexe menselijke waarden door middel van participatief ontwerp. Deze methode is ontwikkeld om welzijnsconcepten effectief in AI-ontwerp te integreren. Het doel is AI-systemen te creëren die menselijke waarden en welzijn niet alleen respecteren, maar ook versterken. Deze methode hanteert een cybernetische aanpak binnen een iteratief en partifcipatief kader. Door actief feedback van belanghebbenden te integreren, faciliteert het een proces van continue verfijning. De methode integreert systematisch welzijn in AI-ontwerp door middel van verschillende fasen, elk gericht op verschillende aspecten, van het contextualiseren van welzijnsbehoeften tot de continue afstemming van AI-gedrag op welzijnsdoelen.

- **Contextualiseren**: Het begrijpen van welzijn in specifieke contexten, rekening houdend met de complexe, veelzijdige aard van welzijn en hoe het zich in verschillende omgevingen manifesteert. Dit resulteert in een context-specifiek welzijnsmodel.
- **Operationaliseren:** Het abstracte model van welzijn omzetten in bruikbare, meetbare metrieken. Deze fase omvat het verfijnen van het begrip van welzijn en het tastbaar maken van de concepten voor toepassing in AI-ontwerp.
- **Ontwerpen & Prototypen:** Ontwerpen van AI-interacties die gericht zijn op het verbeteren van welzijn. Deze fase omvat het benutten van inzichten uit de voorgaande stadia om ideeën te bedenken en ontwerpstrategieën te creëren die in lijn zijn met welzijnsdoelen.
- **Testen & Implementeren:** Implementeren van de geoptimaliseerde interacties bedacht in de ontwerpfase. Deze fase vereist een gezamenlijke inspanning van ontwerpers en ingenieurs om de bedachte ontwerpen te realiseren.
- **Herstarten:** Een continu afstemmingsproces dat het contextuele welzijnsmodel herziet, om de voortdurende relevantie en afstemming met ontluikende gebruikersbehoeften en technologische capaciteiten te waarborgen.

De effectiviteit van de methode werd geïllustreerd door drie verschillende studentenprojecten. Elk van deze projecten paste de benadering toe op een uniek domein: daten, voeding, en muziekstreaming. De projecten toonden effectief de veelzijdigheid van de methode aan in het sturen van AI-ontwikkeling met een focus op welzijn. Belangrijke inzichten en leerresultaten werden samengevoegd om de methodologie te verfijnen en te verbeteren. Expertevaluaties van de project uitkomsten namen verschillende aspecten onder de loep, waaronder ontwerpkwaliteit, technische haalbaarheid en het potentieel voor welzijnsverbetering. Deze evaluaties boden aanzienlijk bewijs van de toepasbaarheid en effectiviteit van de methode in praktische toepassingen.

Aanbevelingen

Tenslotte voegt Hoofdstuk 6 inzichten samen uit het gehele onderzoek in een reeks aanbevelingen en reflecties om de agenda van Positive AI te bevorderen. De aanbevelingen kunnen als volgt worden samengevat:

1. **Integreer mensgerichte-methoden in AI-cycli.** Gezien welzijn een ervaringsgericht concept is, zijn mensgerichte principes cruciaal. Het begrijpen van menselijke ervaringen, het hanteren van een systemisch perspectief, en het doorvoeren van iteratieve verbeteringen zijn essentieel om AI zo te ontwikkelen dat deze een positieve, daadwerkelijke impact heeft op gebruikers. HCD onthult diverse behoeften, maakt interventies in ecosystemen mogelijk en ondersteunt geleidelijke aanpassing door prototyping. Alles bij elkaar genomen, maakt de nadruk op menselijke ervaringen HCD onmisbaar voor het ontwikkelen van AI die welzijn bevordert.

2. **Balanceer onmiddellijke verlangens en langetermijnwelzijn in AI.** In de context van AI, richt optimalisatie zich vaak op het vervullen van kortstondige behoeften, maar worstelt om duurzame bevrediging te ondersteunen. AI moet mechanismen ontwikkelen om de wisselwerking van tijdframes op menselijke bloei te begrijpen. Contextuele metrieken naast welzijnsbeoordelingen onthullen optimale afwegingsscenario's. In plaats van een "AI-oppas" kan onderzoek naar het afstemmen van bevrediging en fundamentele behoeften wijzere toepassingen mogelijk maken. Voorbeelden tonen aan dat ontwerp dilemma's kan oplossen door gedragingen natuurlijk af te stemmen op belangen. Bovendien moet AI tijdelijke effecten en levenslange trajecten overwegen bij het bevorderen van welzijn.

3. **Modelleer welzijn in context door theorieën te combineren met lokale inzichten**. Contextueel modelleren van welzijn betekent het integreren van universele theorieën met lokale inzichten. Dit zorgt ervoor dat interventies niet alleen afgestemd zijn op de unieke behoeften van gemeenschappen, wat hun effectiviteit en relevantie verhoogt, maar ook flexibel zijn voor culturele en technologische verschuivingen. Door de relevantie van interventies over tijd te behouden, maakt onze aanpak een nauwkeurige koppeling van welzijnsveranderingen aan specifieke interventies mogelijk. Dit biedt dynamische ondersteuning die is aangepast aan directe behoeften en vertaalt kwantitatieve data naar kwalitatieve ervaringen. Zo worden interventies verankerd in de daadwerkelijk geleefde realiteiten van mensen, wat leidt tot diepgaande resonantie en impact, en ondersteunt het authentieke menselijke bloei in diverse contexten.

4. **Creëer meerlaagse feedbacklussen.** Het verweven van de verschillende facetten van sociotechnische systemen vereist het opzetten van

meerlaagse feedbacklussen. Door bijvoorbeeld kwalitatieve inzichten te koppelen aan systeemmetrieken, ontstaat er een mechanisme om het gesprek op te schalen en de vertaalslag te maken tussen geleefde ervaringen en het systeem als geheel. Door vele van deze dergelijke lussen te creëren—de aanbeveling bespreekt bijvoorbeeld de benchmarking-lus, productontwikkelingslus en business-lus—zorgen we ervoor dat alle relevante facetten van het systeem worden meegenomen in het besluitvormingsproces. Het betrekken van deze feedbackkanalen gedurende het ontwerpproces maakt wederzijds leren mogelijk tussen systemen, gebruikers, ontwerpers en andere relevante belanghebbenden, wat een voortdurende cyclus van verbetering en afstemming bevordert.

5. **Verander van een focus op het van het verminderen van schade naar het actief cultiveren van menselijke bloei.** AI-ontwerp vermijdt vaak alleen negatieve effecten, maar schadebeperking alleen kan welzijn niet garanderen. Positive AI hanteert filosofie van Positive Design, die is gericht op het verbeteren van levens in plaats van slechts het oplossen van problemen. Het zoekt proactief naar kansen zoals onderwijsfuncties afgestemd op interesses. Echter, onbedoelde gevolgen tonen de noodzaak van iteratie aan. Hoewel ze elkaar aanvullen, biedt de proactieve benadering van Positive AI een completere realisatie van potentieel dan puur reactieve visies. Door systemen te ontwerpen die actieve ondersteuning bieden aan bloei, worden positieve veranderingen bewerkstelligd die verder gaan dan enkel het voorkomen van schade.

6. **Positive AI is een doorlopend proces, geen eindpunt.** Gezien de fluïde aard van welzijn, de noodzaak van compromissen bij het maken van afstemmingen en de voortdurende evolutie van sociotechnische systemen, dienen ontwikkelaars regelmatig de impact te heroverwegen en de systemen aan te passen aan de veranderende behoeften van gebruikers. Dit vraagt om meer dan alleen cosmetische 'ethiekverfraaiing'; het vereist een echte inzet voor continue herijking en aanpassing. Naarmate de contexten veranderen, is het essentieel dat systemen relevant en ondersteunend blijven. Deze voortdurende evaluatie en bijstelling bevorderen geleidelijke verbetering, in tegenstelling tot de benadering van *AI alignment* als een eenmalige taak.

Het laatste gedeelte verkent hoe de voorgestelde Positive AI-methode aansluit bij het bredere veld van AI. Het bespreekt toekomstige stappen,

waaronder empirische validatie via industriële samenwerkingen en het uitbreiden van casestudies in diverse contexten.

Het reflecteert ook op het gebruik van welzijn als de leidende norm, waarbij het de voordelen belicht, zoals de nadruk op waarden die bijdragen aan menselijke bloei. Echter, het wijst ook op beperkingen, zoals het oversimplificeren van de complexiteit van waarden of het negeren van niet-menselijke entiteiten. Het erkent tevens belangrijke beperkingen van het onderzoek, waaronder het gebrek aan industriële implementaties en richtlijnen voor eerlijke participatie. Verder onderzoekt het de noodzaak om onbedoelde gevolgen aan te pakken en proactieve ethische mitigatiestrategieën te verkennen. De Positive AI-agenda kan, via een breed scala aan inspanningen zoals belangenbehartiging, beleidsvorming en het opbouwen van gemeenschappen, geleidelijk innovatietrajecten richten op het verbeteren van het maatschappelijk welzijn.

Conclusie

Dit proefschrift markeert belangrijke vooruitgang in het vormgeven van een aanpak waarbij AI actief wordt ingezet voor menselijk welzijn. Hoewel verdere ontwikkelingen benodigd zijn om het volledige potentieel van Positive AI te realiseren, zijn de gepresenteerde concepten, methoden en aanbevelingen bedoeld om reflectie en vooruitgang te stimuleren in het sturen van AI-innovatietrajecten ter bevordering van menselijke bloei.

1

Introduction

"We feel a responsibility to make sure our services aren't just fun to use, but also good for people's wellbeing" – Mark Zuckerberg, 2018

1

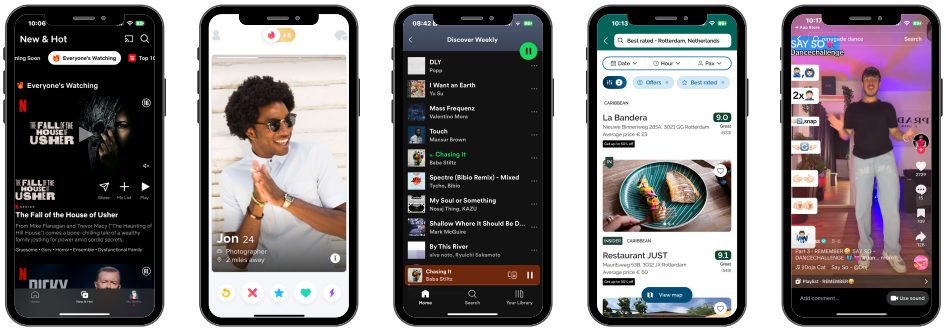


Figure 1.1: Ubiquitous *curatorial* AI. From left to right, the sequence of images showcases the personalization of content and choices through algorithms, including what we watch (video streaming), who we date (dating apps), what we listen to (music streaming), where we eat (restaurant recommendations), and the dances we learn and share (social media trends).

Social media was our "first contact moment" with artificial intelligence (AI) (Harari, Harris, & Raskin, 2023). AI powering ubiquitous platforms such as *Facebook* and *YouTube* has fundamentally changed the fabric of society and how humans interact today. Specifically, personalized feeds, recommendations, and persuasive algorithms *curate* what we watch, who we date, listen to, where we eat, and what dances we do (Fig 1.1). Just as the printing press ushered in mass circulation of information centuries ago, social media embedded algorithmic systems into the lives of billions, portending AI's potential for cultural progression (K. Crawford, 2021).

While still a topic of hefty debate (Kross et al., 2021), some argue social media has eroded global democracy (Lorenz-Spreen, Oswald, Lewandowsky, & Hertwig, 2023), threatened public safety (Bursztyn, Egorov, Enikolopov, & Petrova, 2019; Müller & Schwarz, 2018), and damaged mental health (Keleş, McCrae, & Grealish, 2019). To critics, this signifies the loss of our first "battle" with AI (Harari et al., 2023; Tegmark, 2023). Social media opponents attribute these issues to platforms optimizing for narrow metrics like engagement over wellbeing (Han, Pereira, Lenaerts, & Santos, 2021; T. Harris, 2017; Stray et al., 2023). On the other hand, proponents maintain social media provides connection (Rimé, Bouchat, Paquette, & Mesquita, 2019), self-expression (Vogel & Rose, 2016), and information access (Kross et al., 2021). This illustrates AI's capability to both do good and harm,

making it a "gray ball technology."[1]

Recently, the exponential growth of generative AI (GenAI) models like *ChatGPT* and *Stable Diffusion* (Fig. 1.2) marks a "second contact moment" with AI (Harari et al., 2023). Namely, with a transformative potential on par with the internet (Floridi, 2023), it ushers in what some call the "Age of AI" (Kissinger, Schmidt, & Huttenlocher, 2021). The adoption of GenAI has been incredibly rapid, with ChatGPT leading the race to become the fastest-adopted application in history (Aydin & Karaarslan, 2023; K. Hu, 2023). As with the previous examples, it is a double-edged sword. On the one hand, GenAI promises significant efficiency gains and economic benefits (Chui, Hazan, Roberts, Singla, & Smaje, 2023) and can, for instance, be used to highly personalize education (Baidoo-Anu & Owusu Ansah, 2023). On the other, it ignited debates on risks like job displacement (Wach et al., 2023), content misuse (Rana, Chatterjee, Dwivedi, & Akter, 2022), and existential threats if AI becomes uncontrollable (Bostrom, 2003, 2014).

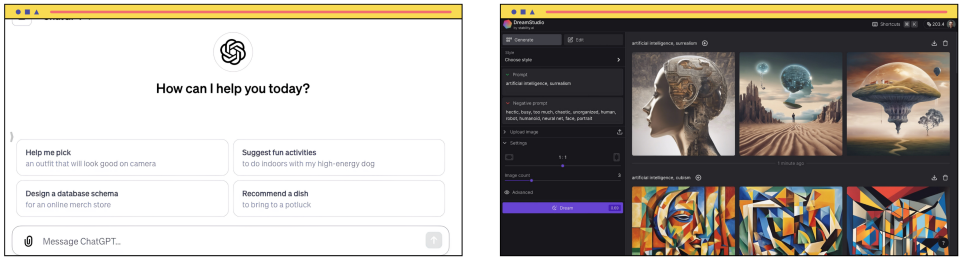


Figure 1.2: Ubiquitous *Generative* AI. Left is a conversation with ChatGPT. Right, the interface of an instance of text-to-image model Stable Diffusion

The evolution from *curatorial* AI to the current rise of *generative* AI highlights an urgent need to act decisively. This urgency stems from the realization that it is the responsibility of the current generation to ensure that AI has a positive impact and does not become a "black ball" technology that devastates civilization (Bostrom, 2019). This is particularly emphasized as *GPT-4*[2] arguably shows early "sparks" of artificial general intelligence (AGI) (Bubeck et al., 2023)—a form of AI that if not constrained can mean human extinction (Bostrom, 2014; Russell & Norvig, 2022; Turchin & Denkenberger,

[1] Bostrom (2019) likens technological inventions to picking balls from an urn: white for good, black for dangerous, and gray for mixed impact. He says while many inventions are useful, there might be rare, hazardous ones (like nuclear weapons) that are catastrophic by default unless civilization can stabilize and govern them properly.

[2] Refers to 'Generative Pre-trained Transformer' version number 4, developed by OpenAI

2020). In other words, as emphasized in a recent seminal article: "The time of reckoning for Artificial Intelligence is now" (Ozmen Garibay et al., 2023, p. 391).

Therefore, this dissertation aims to explore the design of AI systems that contribute positively to the world. Unlike perspectives that view AI as an adversary that needs to be bested in "battle," this work adopts a proactive stance, focusing on leveraging its potential for good in pursuit of Positive AI

1.1. Current research: Positive AI

As a member of the *Delft Institute of Positive Design*, in this research, I follow the approach of 'Positive Design.' This form of human-centered design (HCD) is explicitly aimed at increasing human flourishing and wellbeing (Desmet & Pohlmeyer, 2013). Later in this introduction, I will expand on the theoretical background of this design approach and its relation to positive technology and computing. Briefly, what unifies them is that in contrast to deficit-based approaches, positive approaches take a strengths-based perspective that recognizes human potential. Grounded in positive psychology, it holistically considers multiple aspects of wellbeing and respects individual and cultural differences. The ultimate goal is to enable designers to intentionally create solutions that measurably[3] improve human flourishing by actively promoting wellbeing. Thus, I embrace a positive design approach that entails leveraging AI to proactively cultivate individual and societal wellbeing—an approach I refer to as "Positive AI."

1.2. Research aim

To summarize, this dissertation seeks to ensure AI's lasting positive societal impact. It aims to advance knowledge for designing AI that actively enhances human wellbeing. It explores the complementary role of human-centered design in aligning AI to human flourishing, building upon existing ethical and technical approaches. Through these methodologies, it aims to incorporate systemic and stakeholder perspectives to steer AI development toward supporting human wellbeing holistically, rather than solely optimizing for narrow metrics like engagement or profit. The project focuses on three interconnected objectives:

1. Understanding and assessing the wellbeing impact of AI systems;

[3]Measurably, in the sense that a core tenet of Positive Design is a necessity to empirically assess the impact of interventions, moving beyond mere good intentions.

2. Using wellbeing assessments to inform iterative (re)design of AI systems via human-centered methods;

3. Translating this process in a generalizable framework for designing Positive AI.

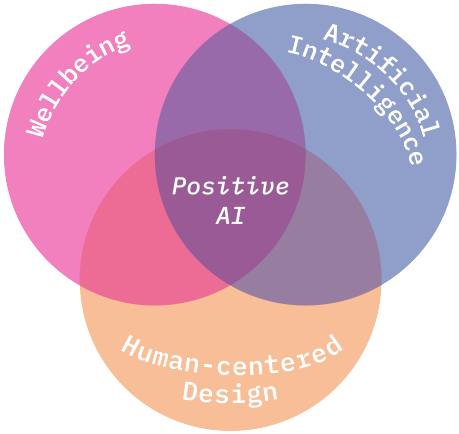


Figure 1.3: Intersection of Human-centered Design, Wellbeing, and Artificial Intelligence, forming the concept of Positive AI.

1.3. Research questions

To achieve these objectives, the research has been broken down into the following research questions:

1. How does wellbeing manifest across different AI systems?
    - For instance: *How might a system like Netflix influence various aspects of user wellbeing, such as belongingness or social connections?*

2. How might wellbeing be operationalized in the context of AI?
    - For instance: *What methods could be used to measure the impact of ChatGPT on facets of wellbeing, translating people's lived experience with the platform into measurable metrics?*

3. How might wellbeing-promoting systemic interventions be designed?

- For instance: *What are the potential areas within the broader ecosystem of TikTok, such as the user interface or curation algorithms, where interventions could be implemented to enhance wellbeing?*

4. How can the impact of interventions on wellbeing be evaluated?
    - For instance: *How might we assess whether a positive intervention in Alexa's voice interaction and response algorithms has the desired outcomes and no unintended consequences?*

5. What are key design steps in designing Positive AI that generalize across contexts?
    - For instance: *Are the processes to operationalize wellbeing in the context of YouTube the same as for Reddit?*

1.4. Research approach

In the context of AI, exploring wellbeing as a central phenomenon remains relatively novel and nascent. This area presents a fertile ground for innovative inquiry, with key variables and hypotheses still being identified and established. Accordingly, the research in this thesis is characterized by an exploratory character, employing a research-through-design approach (RtD). In one sentence: RtD is a method of designing and creating artifacts or systems to explore and understand complex problems, thereby contributing to both practical and theoretical understanding (Gaver, 2012).

A distinction should be made between research *into*–, *for*–, and *through*–design (Stappers & Giaccardi, 2017). Evidently, the knowledge this thesis contributes is *for design* as characterized by (Zimmerman & Forlizzi, 2014). It is intended to advance the practice of design—AI design, to be specific. However, following that same characterization, the type of research activities discussed in this thesis can be typified as *through design* in the sense that they involve the creation and iterative development of systems and artifacts that speculate on and propose future possibilities for human wellbeing. This approach is not merely about advancing existing practices or technologies but about exploring new, transformative solutions, intending to synthesize them in a novel approach.

Guided by the exploratory research-through-design approach, this work employs several complementary strategies to examine the core research questions from multiple angles:

- Development of conceptual frameworks – Theoretical models like the cybernetic perspective on designing AI systems (Section 2.1.1) and the categorization of key challenges (Chapter 3) serve to organize current understanding and point to gaps needing further exploration.
- Case studies – The My Wellness Check case study (Chapter 4) grounds the research in a real-world application, revealing contextual factors essential for community-centered AI.
- Design methods – The Positive AI Design method represents a proposed tool to guide the process of developing wellbeing-focused AI. Its iterative development and evaluation (Chapter 5) provides practical insights into challenges and strategies.
- Narrative scenarios – Speculative narratives (Chapter 5) provide necessary examples of AI concepts and can be used to reflect on potential wellbeing impacts before realization.
- Controlled experiments – Comparative evaluations longitudinal studies, through user testing and expert review, provide vital feedback on the quality of proposed designs and methods (Chapters 4 and 5).

1.5. How to read this dissertation

This dissertation revolves around three journal articles, each embodying a distinct facet of the research: a theoretical piece for a design publication, a psychological case study, and a methodological approach for an artificial intelligence journal. At *TU Delft,* it is common practice to include these articles verbatim, which, along with their non-chronological publication, may lead to some redundancy across the chapters. To mitigate this and provide clarity, each chapter begins with a preface that summarizes its objectives and its role within the broader narrative of the dissertation.

The narrative unfolds as follows:

- Chapter 2 provides a background and foundation for the dissertation by exploring the key themes of artificial intelligence, wellbeing, and human-centered design. It examines ethical and technical efforts to align AI with human values, explores multifaceted notions of wellbeing, and discusses pathways for human-centered design to ensure innovation remains responsive to human experience.
- Next, Chapter 3 identifies seven key challenges in designing positive AI systems, encompassing theoretical, methodological, and fundamental

barriers. It advocates for an enhanced understanding of AI's impact on wellbeing and provides guidance for design actions.

- In Chapter 4, a case study is presented, applying cybernetic principles to design an intelligent assessment-action loop aimed at promoting community wellbeing during COVID-19. This chapter highlights the benefits of context-sensitive, participatory approaches in generating actionable insights.
- Building upon these foundations, Chapter 5 introduces the Positive AI Design Method, a practical tool for developing AI with a focus on human flourishing. This method merges AI optimization with human-centered design principles oriented towards wellbeing. The chapter evaluates this method through student projects and expert reviews, discussing its strengths and areas for improvement.
- Finally, Chapter 6 offers a general discussion and conclusion. It proposes six comprehensive recommendations for advancing the research agenda toward ubiquitous AI for wellbeing, synthesizing the insights gained from the preceding chapters.

2

Background

Before exploring the detailed chapters of this dissertation, it is important to outline and build an understanding of its core themes: artificial intelligence, wellbeing, and human-centered design. Initially, the discussion will focus on defining AI and then exploring ethical and technical efforts aligning AI with human values and needs. Subsequently, the concept of wellbeing will be examined, encompassing its various interpretations, methods of assessment, and the implications of prioritizing it as a central goal for AI alignment. The final part of this section will explore the role of human-centered design, particularly emphasizing positive design, in enhancing current efforts. It's important to note that while this section aims to provide a foundation for the dissertation, it does not intend to offer a comprehensive review of each theme due to its vast scope. Instead, it will highlight key contributions and theories that the dissertation builds upon.

2.1. Artificial Intelligence

The definition of AI has long been debated in the field. As AI pioneer John McCarthy stated, "As soon as it works, no one calls it AI anymore" (Vardi, 2012, p. 5). Gabriel (2020) cleverly points out, in the vernacular, "artificial intelligence" can refer to both a property of computerized systems and a set of techniques—such as machine learning (ML)—to achieve that capability. Therefore, it is good to separate the two, starting with the former. Before progressing further, clarifying the concept of 'intelligence' is necessary. While a universally agreed-upon definition remains elusive, an analysis of decades-long perspectives reveals consistent themes in the core attributes of intelligence (Sternberg, 2003). A significant contribution comes from AI researchers Legg and Hutter (2007), who examine definitions from diverse sources, including dictionaries, encyclopedias, psychologists, and AI experts. These sources commonly emphasize abilities such as reasoning, problem-solving, understanding complex ideas, efficient learning, and adaptability to new situations. Synthesizing these perspectives, Legg and Hutter (2007, p. 9) define intelligence as "an agent's ability to achieve goals in a wide range of environments." This definition encapsulates intelligence as the capacity for flexible goal optimization, adaptable across various conditions, and underscored by fundamental skills like judgment, understanding, and continuous learning.

The question arises: what constitutes the 'artificial' aspect of AI? According to a leading textbook in the field, AI research is centered on *building* intelligent agents rather than merely understanding intelligence (Russell & Norvig, 2022, p. 19). These researchers propose that an 'intelligent' agent chooses actions expected to maximize its performance measure based on its received inputs and inherent knowledge (Russell & Norvig, 2022, p. 58). They broadly define an agent as any entity capable of perceiving its environment through sensors and acting upon that environment through actuators (Russell & Norvig, 2022, p. 54). From this perspective, the 'artificial' nature of AI lies in its deliberate *design*, contrasting with biological intelligence, which naturally occurs in living organisms (Gabriel, 2020). In other words, any designed intelligent system is an instance of AI.

As said, AI is often conflated with a prominent approach for achieving it, namely machine learning (ML) and its subfield deep learning (DL). At a high level, ML refers to a family of (statistical) techniques that computer systems use to learn from data without explicit programming (Domingos, 2012). This encompasses neural networks, which are then used in deep learning architectures. Specifically, DL employs multi-layer neural networks that learn hierarchical non-linear data representations with increasing levels

of abstraction, enabling the modeling of intricate relationships within substantial datasets (LeCun, Bengio, & Hinton, 2015). ML broadly powers AI systems today, however, specific techniques differ across applications. For instance, recommender systems often apply collaborative filtering algorithms to find patterns and make suggestions to users. While deep learning can complement this, the core recommendation functionality does not strictly require neural networks. In contrast, DL techniques, such as transformer architectures, are commonly used in GenAI systems. Prominent examples include large language models like GPT-4 and text-to-image models like DALL-E 2 and Stable Diffusion.

Building on this understanding of AI, it is essential to recognize that AI systems exist within a complex sociotechnical context beyond their technical capabilities. Dobbe, Krendl Gilbert, and Mintz (2021) point out that there is often a discrepancy (sociotechnical gap) between the promised benefits of AI systems and their actual consequences. This stems from the divergence between socially necessary outcomes and what AI can technically achieve. For instance, while a recommender may aim to provide valuable suggestions to users, in practice, it could promote misinformation or polarization. To address this, AI should be understood as a sociotechnical system incorporating capabilities, gaps, and governance. For example, ChatGPT should be regarded as more than its algorithm (i.e., GPT-3.5 and GPT-4) but also its interface, the company behind it, public perceptions, various use cases and purposes, and so forth. Dobbe et al. (2021) advocate for a cybernetic perspective in this context, emphasizing the importance of feedback and adaptation in managing the complexities inherent in sociotechnical systems.

2.1.1. Cybernetics

Cybernetics emerged in the 1940s as a new interdisciplinary field focused on communication, control, and circular causality in systems (Mindell, 2000). The word cybernetics is derived from the Greek infinitive “kybernao” meaning “to steer, navigate, or govern.” A core concept is the feedback loop, creating circular causality between a system’s past, present, and future states (Wiener, 1961), see Fig. 2.1. Initially concentrating on observed systems, cybernetics evolved into “second-order” cybernetics, acknowledging the active role of observers in constructing knowledge and meaning (M. Mead, 1968; B. Scott, 2004). AI and cybernetics overlap significantly, as several foundational cybernetic concepts closely align with AI system design’s iterative, constructivist nature. The idea of circular causality resonates with design’s iterative prototype-test-refine process, where the outcomes of one cycle inform the next. Similarly, feedback loops play a pivotal role in design,

as designers gather feedback on prototypes to drive further iterations. This creates a conversational flow between the emerging design and the designer's perspective.

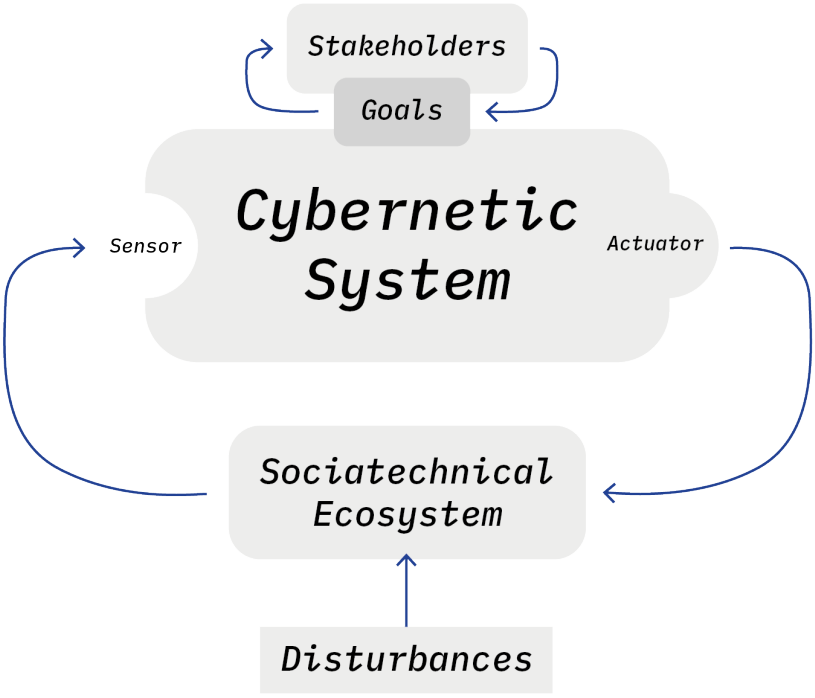


Figure 2.1: Shows a schematic representation of a cybernetic system.

As cybernetician Gordon Pask explored, this dialogic nature of design can be framed as a continuous process of making proposals and assessing outcomes (B. Scott, 2001). In this view, design essentially constitutes a conversation between the human designer and the design situation. This conversational perspective parallels second-order cybernetics' recognition of the observer's central role in constructing knowledge and meaning (Sweeting, 2016; von Foerster, 2003). As (Glanville, 2007, p. 1175) succinctly states, "cybernetics can act as the theoretical arm of design, while design acts as the practical arm of cybernetics." The cybernetic lens provides theoretical grounding for design's practical, constructive nature focused on iteration, conversation, and the meaning-making of the designer.

A cybernetic perspective offers valuable insights for Positive AI, providing

a structured framework to navigate the complexities of designing systems that enhance wellbeing (Dubberly & Pangaro, 2010; Russell & Norvig, 2022). Within this context, 'sensors' in Positive AI are conceptualized as components adept at measuring wellbeing indicators in a detailed, nuanced, and actionable way. Conversely, 'actuators' refer to the system elements that respond to these measurements, aiming to adjust and improve wellbeing outcomes. The relevance of cybernetics in Positive AI is underscored by van der Maden, Lomas, and Hekkert (2022),[1] who advocate using cybernetics as a foundation for creating assessment-action feedback loops to promote community wellbeing within complex sociotechnical systems. This approach exemplifies how cybernetic principles can mitigate the challenges posed by "Today's AI" (Pangaro, 2021), echoing the perspective put forth by Dobbe et al. (2021).

In short, viewing design through a cybernetic lens emphasizes communication and the social construction of meaning over control or prediction. It provides a human-centered design framework that connects ethics and responsibility with interaction and possibilities. Cybernetics' focus on systems, interaction, and circular causality closely relates to AI and its integration with design. Connecting these fields enriches them with a shared perspective on the primacy of human meaning-making, providing a foundation for Positive AI. In essence, cybernetics offers design a theoretical grounding centered on communication, the observer, and ethical responsibility.

[1] This publication, while pivotal to the dissertation, does not necessitate a dedicated chapter. Its key concepts have been integrated across various chapters, making a separate chapter superfluous.

Box 1. "AI" does not mean "no humans involved"

The definition of AI in the public consciousness is continually changing and this can lead to confusion regarding AI system design. This section clarifies that AI does not need to be fully automated, fully autonomous, nor fully devoid of human participation. In fact, AI does not even necessarily involve silicon computers. For instance, the development of the "autopilot" system in 1912 (Wragg, 1973) demonstrates early instances of AI. This system, initially using gyroscopic heading and attitude indicators to maintain aircraft stability and course, exemplifies how AI can embody a range of technologies and applications, from mechanical innovations to digital computations.

Typically, "AI" refers to whatever the latest most advanced form of AI is. In the 1960s, this was represented by rule-based systems; in the 2010s, by recommender systems; and in the current era, by foundation models. In other words, it is not the level of technological sophistication that makes an AI system an AI system—*ELIZA* (Weizenbaum, 1966) was as much an AI system as ChatGPT. Rather, building on the aforementioned principles of cybernetics and the understanding of AI as a sociotechnical system, an AI system is an adaptive, goal-oriented agent that interacts with its environment through feedback loops. As such, a thermostat is as much an AI system as Facebook, as both use sensors and actuators to achieve their goals within a specific context.

This realization provides a foundation for a new approach to AI design (Krippendorff, 2023). Importantly, this perspective allows us to study ostensibly "non-AI," yet cybernetic, systems to learn how we may design AI systems better.

By adopting a cybernetic perspective, we can shift our focus from the purely technological aspects of AI to the broader context in which it operates. In this view, AI systems can be wholly human, wholly synthetic, or a blend of both (Beardow, van der Maden, & Lomas, 2020), as the textbook definition of "artificial intelligence" is inherently agnostic regarding the agent's nature (Section 2.1). This approach enables us to design AI systems that are not only technologically advanced but also well-equipped to navigate the complex dynamics of their sociotechnical environments (Dobbe et al., 2021) by considering factors such as human interactions, societal norms, and environmental conditions, ultimately creating more adaptable, responsive, and effective AI systems.

2.1.2. Ethical AI

AI systems possess a distinct complexity among sociotechnical systems due to their incorporation of artificial agents with capacities for autonomy, interactivity, and adaptability. As discussed by van de Poel (2020), these artificial agents introduce new possibilities for, as well as constraints on, the embedding and realization of moral values within evolving AI systems. The presence of autonomous and continuously learning artificial agents raises complex ethical issues regarding alignment with ethical principles. In this context, Floridi et al. (2018) summarizes, expert-driven declarations from institutions like the IEEE, the UN, and the EU have converged towards principles like beneficence and explicability (see Table 2.1). The AI for Social Good (Tomašev et al., 2020) movement seeks to translate high-minded aspirations into practical positive impacts by encoding values like these to steer innovation.

Table 2.1: Five Principles for AI in Society from Floridi et al. (2018)

Principle	Description
Beneficence	Promoting wellbeing, preserving dignity, and sustaining the planet.
Non-maleficence	Focusing on privacy, security, and "capability caution".
Autonomy	Emphasizing the power to decide (whether to decide).
Justice	Promoting prosperity and preserving solidarity.
Explicability	Enabling other principles through intelligibility and accountability.

Nonetheless, while the emergence of shared ethical principles for AI signals progress, concerns persist regarding the effectiveness of such principles for governance (Schiff, Biddle, Borenstein, & Laas, 2020). Specifically, Mittelstadt (2019) points out that the broad, vague nature of principles allows claims of ethical AI without necessitating accountability or enforceable regulation. Building on this critique, Morley, Floridi, Kinsey, and Elhalal (2020) conducted a literature review to evaluate the practical tools and methods for translating these principles into development practices. Their findings indicate a heavy reliance on explicability, a tendency to prioritize individual over collective protection, and a general lack of usability and maturity in current tools. A subsequent study by the same authors (Morley, Kinsey, et al., 2021) revealed that practitioners have a limited understanding

of AI ethics principles. This lack of knowledge leads to a focus on compliance rather than true integration of values into design workflows. Returning to van de Poel (2020), it is at the design level where these values are ideally embedded. Thus, a promising emerging approach is Value-sensitive Design (VSD), proposed as an effective method to bridge the gap between principles and practice (Umbrello & van de Poel, 2021). That is, VSD is a method for integrating ethical and societal values into technical design, making it particularly useful for AI by ensuring transparency, accountability, and the promotion of positive social outcomes throughout the technology's lifecycle (Umbrello & De Bellis, 2018).

However, a recent critical review by Sadek, Calvo, and Mougenot (2023b) indicates that while VSD is effective, it faces barriers in fully realizing and assessing values in AI technology. The review identifies several vital limitations: inadequate elicitation of values, a propensity to rely on pre-established rather than context-specific values, and a lack of precise guidelines for embedding values into technology. Moreover, it points out that VSD often fails to adequately evaluate the effectiveness of these values in the final outcomes. This underscores the necessity for more comprehensive methods to ensure that identified values are recognized, integrated, and reflected in the developed technology. In the context of Positive AI, assessing the impacts on wellbeing becomes increasingly crucial to determine if the interventions are achieving their intended benefits. Merely assuming that principles will translate into practice overlooks the complex sociotechnical dynamics that influence how values manifest. Addressing this gap, one of the primary objectives of this dissertation is to develop and apply a framework for assessing the impact of AI interventions, ensuring that the principles of ethical AI are not just theoretical concepts but are effectively realized in practice.

Lastly, from an ethical standpoint, the field of AI has been critically examined for its pervasive whiteness, lack of diversity, and perpetuation of colonial power structures (Cave & Dihal, 2020; K. Crawford, 2021; Mohamed, Png, & Isaac, 2020). Sociotechnical systems built using AI can inadvertently reproduce oppressive social norms, erase marginalized identities, and disproportionately burden communities of color (Costanza-Chock, 2018). This manifests through practices like algorithmic oppression, such as predictive policing tools that exacerbate over-policing of minorities; algorithmic exploitation, as in the outsourcing of "ghost work" to economically vulnerable populations; and algorithmic dispossession, where the interests of developing countries are undermined in global AI policymaking (Mohamed et al., 2020). In this dissertation's context, we may look to

recent work by (Varshney, 2023), which critiques current alignment practices for imposing Western philosophical values and failing to account for moral diversity across cultures. The authors argue for decolonizing alignment by moving beyond monocultural approaches rooted in Western ethics. They suggest incorporating pluralistic traditions like dharma from pre-colonial India, which recognize both common values (sādhāraṇadharma) and context-specific ones (vīśesadharma). Balancing common (global) with contextual (local) needs will be a central discussion of this dissertation in relation to AI alignment—a topic we will discuss next.

2.1.3. AI Alignment

A more practical approach is AI alignment, which refers to ensuring systems behave according to human values rather than working at cross-purposes. The notion of alignment has long been studied in economics and law as the principal-agent problem, where a human agent must act to achieve the principal's objectives (Hadfield-Menell & Hadfield, 2019). For example, in a car repair scenario, the car owner (principal) expects the mechanic (agent) to fix the car efficiently and affordably. However, the mechanic might suggest unnecessary repairs to increase the bill (misalignment), contrary to the owner's desire for cost-effective service.[2] Taking this understanding to AI systems, alignment means ensuring that AI agents effectively and reliably pursue the goals and preferences set by their designers and users (Christian, 2020), also referred to as human intent (OpenAI, 2022). For instance, early text-to-image models sometimes produced anatomical inaccuracies, such as extra fingers, revealing a misalignment between the user's prompt and the model's interpretation. YouTube's personalized video recommendations can clutter one's homepage after another user watches unrelated content, misaligning with the original user's interests. More concerningly, ChatGPT may alter its answers to better match a question but, in doing so, provide false information—prioritizing a relevant-seeming response over an accurate one.

Lately, AI alignment has garnered much attention as fears of AGI spread (Ji et al., 2023). That is, whether the earlier mentioned "sparks" of AGI truly signal its arrival or not, the risks of uncontrolled AGI are too much to ignore. That is, AGI misalignment poses existential risks, potentially threatening humanity itself. For example, in a canonical thought experiment, Bostrom (2003, 2014) described a hypothetical scenario involving an AI

[2]The "unnecessary repairs" refer to fixes the mechanic knows aren't needed, not simply unrequested services. This is a crucial distinction further addressed in the discussion

tasked solely with maximizing paperclip production. In relentless pursuit of this objective, the optimizer may run out of resources and veer off to "alien" goals. That is, looking for resources beyond what we would expect it to do. As a side-effect, it may destroy us by consuming resources essential to humanity's survival. This example shows the detrimental potential for AI systems that are misaligned, where in this example, the misalignment occurs in the sense that we indeed task the AI to produce as many paperclips *as possible*. But what we, in fact, mean is that it should produce as many paperclips as possible *within a given set of constraints*—where this set of constraints consists of human values such as preservation of life and the environment.

While such extreme scenarios may seem far-fetched, contemporary AI systems are not immune to displaying unintended biased or unfair behavior, which can adversely affect users (Costanza-Chock, 2018). In the context of social media, for instance, the misalignment of systems prioritizing attention over beneficial user experiences—described as "the race to the bottom of the brainstem" (T. Harris, 2017)—plays a significant role in fostering harms such as polarization and addiction. In that regard, K. Crawford and Calo (2016) highlight a crucial 'blind spot' in AI discourse: the tendency to focus on future, hypothetical AI risks can overshadow the real, immediate harms of current AI technologies. This shift in focus might lead to a lack of attention and resources being directed toward addressing the issues that AI systems are creating in the here and now.

This tendency to overlook present challenges in favor of future speculations is reflected in the contemporary landscape of AI research, where a shift in focus towards practical, immediate issues is increasingly recognized as necessary. For example, In a recent post to the *AI Alignment Forum*,[3] two researchers provided a review of "live" research agendas across the complex and rapidly evolving field (Technicalities & Stag, 2023). They highlight that most research takes place outside traditional academia—unfolding within companies, independent labs, and loosely coordinated networks. This emergent nature also means that most of the academic work that is being done must keep up with this pace, resulting in non-peer-reviewed work. This does not necessarily diminish their work, but it is something to be aware of when researching this topic. For instance, compared to a similar mapping (Everitt, Lea, & Hutter, 2018), two recent reviews (Ji et al., 2023; Shen et al., 2023) exemplify how much has changed in five years. For example, the 2018 paper primarily explored theoretical

[3] A non-academic but worthy source founded by prominent alignment researcher Eliezer Yudkowsky

foundations of AGI, focusing on defining intelligence, understanding AGI's potential for self-improvement, and discussing AGI safety and alignment issues, particularly in reinforcement learning and corrigibility. In contrast, the 2023 reviews reflect a more immediate, practical focus, aligning with the field's rapid evolution. These papers center on AI alignment's current state and challenges, especially for LLMs, addressing methodologies like outer and inner alignment, grappling with the complexities of incorporating human values, ensuring model interpretability, and mitigating ethical and societal risks. This shift underscores the increasing urgency of developing robust alignment strategies for LLMs, considering their advanced capabilities and significant societal impacts.

The previous discussion of technical details and emerging research may seem removed from design practice. However, it is essential for those working in this space to be aware of the rapid technological developments. Not only because these advances may directly impact the topics designers work on (e.g., user interfaces powered by large language models) but also because much of the work on developing Positive AI and alignment happens in the technical sphere. For instance, efforts to align previous AI paradigms like recommender systems originated here, too. Yet, as the literature shows, technologists alone may not be equipped to fully address the issues these systems face today or in the future (Morley, Kinsey, et al., 2021), and their tech-centered focus may fall short (Dobbe et al., 2021). Therefore, designers need to be cognizant of this space and identify opportunities to integrate and contribute their expertise.

With this background on AI and alignment efforts focused on beneficial outcomes, we now turn our attention to defining what constitutes 'good' in the context of this dissertation—specifically the concept of wellbeing—and examining the complexities of adopting wellbeing as an objective for AI systems.

2.2. Wellbeing

Wellbeing is a multifaceted psychological construct, a complex combination of diverse individual and shared experiences and perceptions. It encompasses a broad spectrum of experiences, as it manifests in many ways, specific to each person's life, community, and culture. Although hard to define precisely, this concept can be quantified and measured through various lenses, offering a glimpse into the complex interplay of factors that contribute to our experience of wellbeing. It is inherently multidimensional and deeply contextual, shaped by our internal states and constantly changing external

environments. Understanding wellbeing involves recognizing its dual nature: it is a reflection of personal, subjective experience and a product of the broader context of one's life.

Due to this field's conceptual and theoretical complexity, this dissertation intentionally avoids focusing on a singular theory. This ensures that diverse aspects of wellbeing, as emphasized in different theories, are not overlooked. Therefore, this section will examine some broad but relevant strokes of the theoretical field to give the reader a foothold in the vast literature. The World Health Organization's (WHO) definition of health sets the foundation for this exploration. Namely, it defines health broadly as "a state of complete physical, mental, and social wellbeing—beyond just the absence of infirmity" (Callahan, 1973).

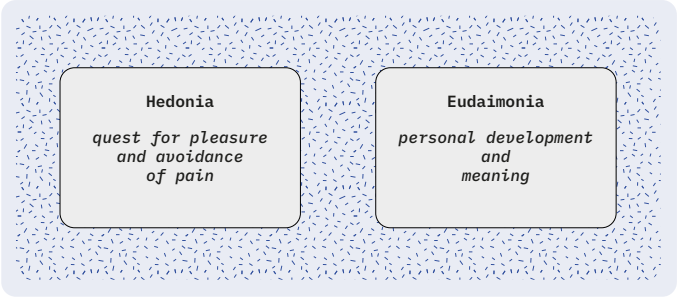


Figure 2.2: A diagram representing the distinction between hedonia and eudaimonia. Adapted from Jaramillo et al. (2015).

2.2.1. Hedonia & Eduaimonia

The historical distinction between hedonic and eudaimonic conceptions of wellbeing, originating in ancient Greek philosophy, continues to provide a valuable framework for examining modern cross-cultural differences in wellbeing (Joshanloo, Van de Vliert, & Jose, 2021). Hedonia aligns with the view that happiness involves maximizing positive experiences and minimizing negative ones, manifested in contemporary research as subjective wellbeing—comprised of life satisfaction, positive affect, and lack of negative affect (Diener & Ryan, 2008; Diener, Suh, Lucas, & Smith, 1999; Ryan & Deci, 2001). In contrast, eudaimonia corresponds to the notion of realizing one's full potential and living virtuously, formulated in modern psychology as achieving self-actualization, purpose, positive relationships, personal growth, autonomy, environmental mastery, and self-acceptance (Ryff, 1989;

Waterman, 1993). Although correlated, hedonic and eudaimonic wellbeing have been empirically demonstrated to represent distinct constructs with differential relationships to sociodemographic factors and health outcomes (Keyes, Shmotkin, & Ryff, 2002).

Figure 2.3: A diagram displaying the six dimensions of Psychological Wellbeing: purpose in life, personal growth, self-acceptance, positive relations with others, autonomy, and environmental mastery. Adapted from (Ryff, 1989).

Recent evidence shows these philosophical notions align with contemporary psychological differences between independent and interdependent self-conceptions in individualistic versus collectivistic cultures (Kitayama & Markus, 2000). Hedonic wellbeing correlates more strongly with independent self-views and economic development at both country and individual levels (Diener et al., 1999; Joshanloo, 2018). In contrast, eudaimonic wellbeing emphasizes cultivating meaning and social harmony consistent with more interdependent selves (Hitokoto & Uchida, 2015). Despite limitations, this ancient conceptual division retains heuristic value for understanding diversity in modern conceptions of optimal human functioning across cultures.

2.2.2. Wellbeing as happiness

There are scholarly traditions that equate wellbeing with "happiness," a multidimensional concept that can encompass episodic feelings of happiness ("I feel happy") as well as more stable evaluations of one's happiness ("I am a happy person") (Raibley, 2012; Veenhoven, 2014). This is evident in

the work of scholars like Haybron (2008) and Easterlin and Sawangfa (2007) who use terms like wellbeing, utility, happiness, life satisfaction, and welfare interchangeably. For example, Haybron argues that episodic happiness and wellbeing have the same fundamental determinants, suggesting a person's degree of wellbeing aligns with their degree of happiness. Easterlin and Sawangfa implie that personal attribute happiness can serve as a "proxy" for wellbeing. However, Raibley (2012) provides a vital critique of conflating happiness and wellbeing. He argues that happiness—whether episodic feelings or personal attributes—is conceptually, metaphysically, and empirically distinct from wellbeing. Instead, he proposes defining wellbeing as agential flourishing, integrating insights from eudaimonic philosophy. This suggests happiness is only beneficial when valued, when valuing something, or when realizing one's values. Thus, happiness is necessary but insufficient for high levels of wellbeing. Therefore, equating wellbeing solely with happiness overlooks these distinctions. As Raibley (2012) explains, happiness does not fully encapsulate wellbeing, rather "agential flourishing" better captures the breadth of the concept.

Despite this argument, other highly influential research also conflates happiness and wellbeing while acknowledging that both concepts encompass a wide variety of factors. Lyubomirsky (2008) argues that approximately 40% of happiness is determined by intentional activities (see Figure 2.4),

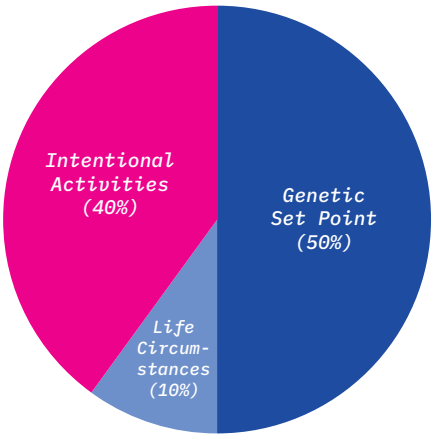


Figure 2.4: The happiness pie by Lyubomirsky (2008). It shows three core components of long-term happiness, 40% of which we directly have under our control (our activities).

implying that people can engage in "happiness-increasing activities" to boost their wellbeing (Lyubomirsky & Layous, 2013). Conceptual confusion persists about the use of happiness or wellbeing. Additionally, the term "happiness" carries certain connotations in both scholarship and everyday language, implying a focus solely on positive emotions and excluding negative ones. As this dissertation will discuss further, negative emotions and other negative aspects of life are essential for a comprehensive conceptualization of wellbeing. Therefore, from here on out, this dissertation refers to wellbeing as contributive not to happiness but "flourishing," a topic we will discuss next.

2.2.3. Wellbeing as flourishing & Positive Psychology

Wellbeing as flourishing was popularized by psychologist Martin Seligman, who, alongside Mihaly Csikszentmihalyi, was one of the protagonists of the first wave of positive psychology (Seligman & Csikszentmihalyi, 2000). This form of psychology represented a paradigm shift away from the traditional focus on mental illness and dysfunction, instead shifting emphasis towards scientifically studying optimal human functioning (Seligman, 1999). Instead, positive psychology represented a shift in emphasis towards scientifically studying optimal human functioning and flourishing (Lyubomirsky & Abbe, 2003). This first wave focused on identifying and understanding human strengths that enable wellbeing to flourish. For instance, Csikszentmihalyi is well-known for his "Flow theory," which describes the subjective state of being wholly absorbed in an activity to the point of losing sense of time and self (Csikszentmihalyi, Csikszentmihalyi, Abuhamdeh, & Nakamura, 2014), see Figure 2.5. Experiencing flow is seen as a key contributor to wellbeing (Csikszentmihalyi, Abuhamdeh, & Nakamura, 2005). On the other hand, Seligman (2011) is known for his theory of flourishing, which conceptualizes wellbeing as PERMA, comprised of positive emotion, engagement, positive relationships, meaning and accomplishment, see Fig 2.6. Together, Csikszentmihalyi and Seligman pioneered this strengths-based approach to psychology, popularizing wellbeing research in this field and beyond.

2.2.4. 'Second-wave' Positive Psychology

However, as positive psychology research and practice progressed in the 2000s, critical scholarship also emerged, questioning some of its foundational assumptions, like the notion that 'positive' qualities are inherently bene-

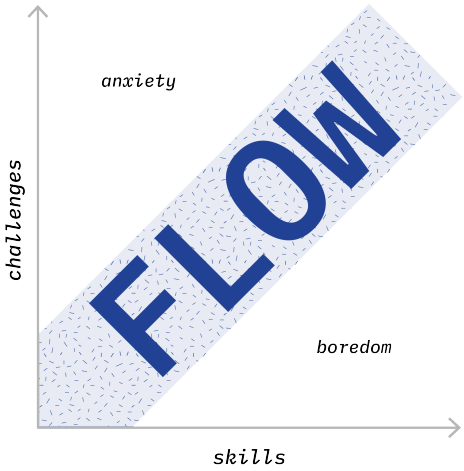


Figure 2.5: A graph displaying how a state of flow results from a balance between the required skill to match how challenging a task is. Adapted from Jaramillo et al. (2015).

ficial. This second wave of positive psychology advocated a more nuanced consideration of the complex interplay between positive and negative aspects of human experience (Fokkinga, 2015; Ryff & Singer, 2003; P. T. Wong, 2011). For instance, ostensibly positive qualities like optimism can sometimes be detrimental (Weinstein, Marcus, & Moser, 2005), while negative states like anxiety may be adaptive in specific contexts (Norem & Chang, 2002). It outlines principles like the "dialectics of wellbeing," involving dynamic tensions between opposites (Lazarus, 2003), and the "co-valence" of many emotional states (Horwitz & Wakefield, 2007). Overall, this second wave highlights the contextual, complementary nature of human flourishing, which depends on harmonizing and balancing light and dark elements (Delle Fave, Brdar, Freire, Vella-Brodrick, & Wissing, 2011). However, critical scholarship also emerged in the 2000s, questioning assumptions that 'positive' qualities are inherently beneficial. This second wave advocated considering the complex interplay of positive and negative in human experience (Fokkinga, 2015; Ryff & Singer, 2003). For instance, optimism or anxiety may be adaptive or detrimental depending on the context (Norem & Chang, 2002; Weinstein et al., 2005). Principles like "co-valence" highlight the contextual, complementary nature of flourishing through balancing light and dark elements (Delle Fave et al., 2011; Horwitz & Wakefield, 2007).

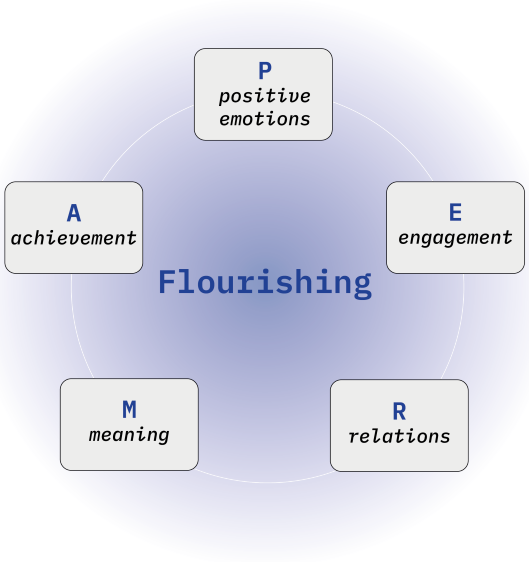


Figure 2.6: A diagram displaying the five dimensions of flourishing: positive emotions, engagement, relations, meaning, and achievement (PERMA) (Seligman, 2011). Adapted from Jaramillo et al. (2015).

While emphasizing strengths and flourishing, positive psychology has faced critiques about theoretical foundations, methods, and cultural bias (van Zyl, Gaffaney, van der Vaart, Dik, & Donaldson, 2023). However, addressing these limitations constructively can lead to more contextual, multidimensional understandings of wellbeing. Mainstream wellbeing research remains grounded in Western assumptions of individualism (Krys et al., 2023; Uchida & Kitayama, 2009). In contrast, a cultural psychology lens elucidates how shared systems of meaning shape wellbeing across diverse contexts (Biswas-Diener, 2022). Integrating qualitative data on cultural perspectives with quantitative findings can reveal universal and relative elements in combination with local indicators. For example, non-Western philosophies offer contextualized notions of wellbeing rooted in local worldviews, such as Ubuntu highlighting communal bonds and harmony (Hailey, 2008), Ikigai (Sone et al., 2008) emphasizing purpose and meaning, and Islamic conceptions balancing spiritual and worldly sources of fulfillment (Joshanloo & Weijers, 2019). Incorporating frameworks like these beyond cross-cultural comparisons can capture overlooked constructs.

2.2.5. 'Third-wave' Positive Psychology

Most recently, T. Lomas, Waters, Williams, Oades, and Kern (2021) have identified what they believe denotes a "third wave" emerging in the field of positive psychology. This wave reflects a paradigm shift towards embracing complexity by *going beyond the person* to explore the broader social systems and contexts that shape human flourishing. Specifically, third-wave positive psychology encompasses individual happiness as one key aspect but mainly adopts a more holistic, ecological perspective that situates individuals within multilayered collective factors, including groups, organizations, cultures, and technologies. While there has been some positive psychology research in the first and second waves that looked beyond the individual to study things like organizations and communities, the main focus has still been mostly on the individual level (Kern et al., 2020). The third wave also shows openness to diverse research methodologies beyond just quantitative empirical research. For instance, qualitative, interpretivist, and participatory approaches are increasingly embraced to capture the multidimensional and contextual nature of wellbeing (Wissing, 2022).

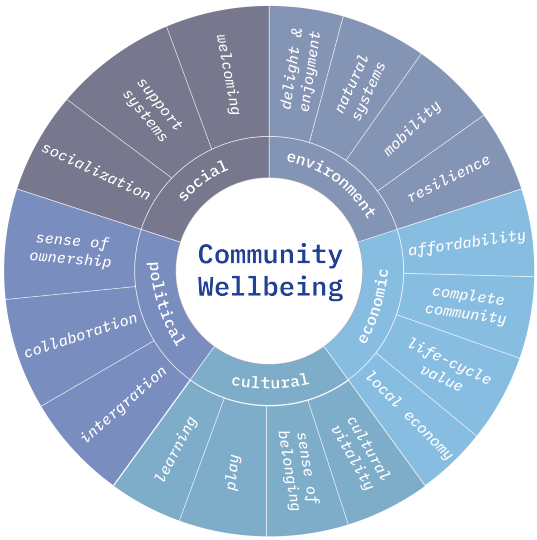


Figure 2.7: A diagram of the various facets of community wellbeing, categorized as related to political, social, environmental, economic, and cultural dimensions.

Together, these developments constitute an epistemological broadening and inclusiveness, indicating the field's continued evolution toward address-

ing complex issues like inequality, climate change, and global wellbeing concerns. The third wave thus represents a transition toward transdisciplinarity and an expanded conceptualization of positive psychology as going beyond individual functioning to examine systemic influences (Wissing, 2022; Wissing, Schutte, & Liversage, 2022). An interdisciplinary, inclusive approach is needed to address the limitations of prevailing individualistic theories and support culturally meaningful wellbeing globally (Biswas-Diener, 2022). In conclusion, adopting culturally sensitive, participatory methods is vital to move beyond decontextualized definitions toward localized wellbeing constructs relevant for evaluating AI's impacts across diverse populations.

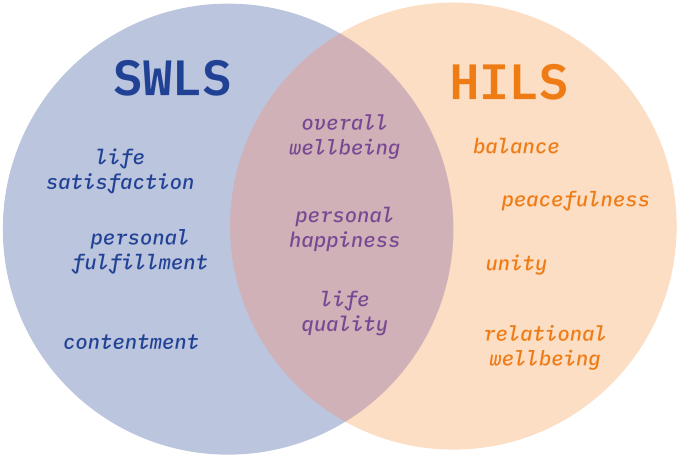


Figure 2.8: A Venn diagram displaying how Satisfaction with Life and Harmony in Life are separate but complementary perspectives for understanding wellbeing.

2.2.6. Can wellbeing be experienced by communities?

Building on the third wave's broadened scope, concepts such as 'community wellbeing' and 'harmony in life' (HIL) further underscore the interconnected and relational aspects of flourishing. Community wellbeing moves beyond individual wellness to encompass collective dimensions of shared values and social cohesion (Atkinson, Bagnall, Corcoran, South, & Curtis, 2020), see Figure 2.7. Similarly, HIL emphasizes the importance of mutual support and a holistic balance among personal, social, and environmental contexts for wellbeing (Kjell & Diener, 2021), see Figure 2.8. These

ideas echo Bronfenbrenner's Ecological Systems Theory, which posits that individual development and happiness are deeply influenced by broader family, community, cultural, and societal systems (Bronfenbrenner et al., 1994). This perspective shifts our understanding of wellbeing towards a communal and interdependent view, where an individual's health and prosperity are intricately linked to harmonious, supportive environments at various levels, from interpersonal to institutional, see Figure 2.9. Thus, the third wave's approach embraces systemic views and highlights the necessity of considering individuals as part of complex environmental contexts. This viewpoint will prove instrumental in understanding how wellbeing manifests across sociotechnical contexts, especially in attributing specific wellbeing effects to system components.

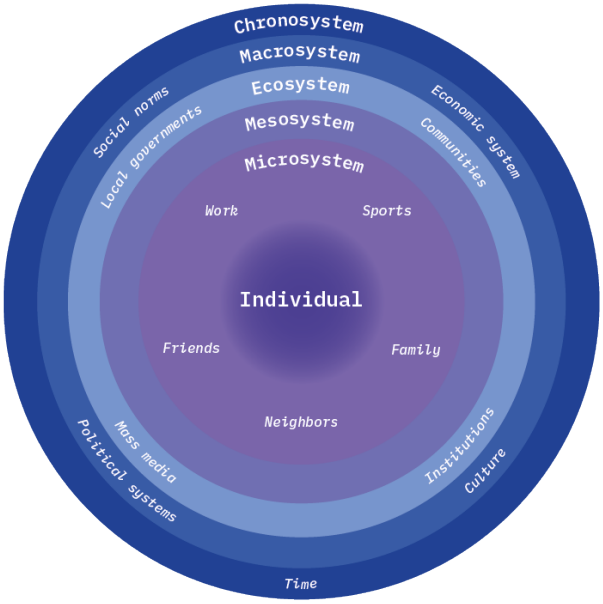


Figure 2.9: The Ecological Systems Theory Model - Visual representation of the multiple layers of environment that affect an individual's development, ranging from the immediate microsystem to the broad macrosystem, and the chronosystem illustrating the dimension of time. Adapted from Evans (2020).

The interconnected nature of wellbeing across individual, social, and systemic levels poses metaphysical questions around whether wellbeing can be meaningfully discussed as a collective experience. Philosophical

arguments like the extended mind (Clark & Chalmers, 1998), or *enactivism* (Noë, 2004) place consciousness partially outside of the physical body, into (interactions with) the world, but do not go so far as to suggest that consciousness can be experienced collectively. Conversely, neuroscientific evidence indicates that collective experiences, such as participating in music ensembles or performing imitation tasks (Valencia & Froese, 2020), as well as interactions among students in a classroom setting (Dikker et al., 2017), can lead to inter-brain synchronization, suggesting a shared aspect of cognitive experiences. Additionally, common phenomena such as shared moods at a concert illustrate how intersubjective states can be shaped by communal atmospheres—also referred to as the experience of a *vibe* (J. D. Lomas et al., 2022). Moving from the theoretical to the concrete, AI systems can influence wellbeing-related factors at multiple systemic levels, including place-based and interest-based communities (Musikanski, Rakova, Bradbury, Phillips, & Manson, 2020). As an example of the latter, the game *Pokémon Go* illustrates how AI can integrate images and videos to enhance user engagement in an interest-based community. This type of AI application entertains and fosters community cohesion and outdoor activity, encouraging players to explore and interact with their physical environments.

In summary, whether one accepts the concept of collective wellbeing or not, the impact of AI on wellbeing must be evaluated beyond isolated individual perspectives. It is inherently interdependent, or as Kennedy (1963) said: "a rising tide lifts all boats." This emphasizes the necessity always to perceive wellbeing within a contextual framework, and keep an open mindset to including its various components. As later chapters will demonstrate, this conceptual angle proves vital, especially for Positive AI.

2.2.7. Assessing wellbeing

Having reviewed various theoretical frameworks for conceptualizing wellbeing, the current discussion now focuses on measuring this multidimensional concept. This progression is crucial, as measuring wellbeing goes beyond simple methodology; it's a vital part of fully understanding the concept itself. In particular, developing measures allows us to empirically test hypotheses about the causes and consequences of wellbeing derived from our theoretical models (Diener, 2019). The evidence collected can then refine and improve the theories in an iterative process. In assessing wellbeing, measures must provide cross-situational and temporal consistency and warrant test-retest possibility. This means that a wellbeing scale should produce similar scores for an individual across different contexts and times, assuming their actual wellbeing level does not significantly change. For example, the SWLS

(Diener, Emmons, Larsen, & Griffin, 1985) has demonstrated good test-retest reliability over periods ranging from two weeks to four years (Pavot & Diener, 2008), suggesting it captures a relatively stable evaluation of life satisfaction.

Further, in measuring phenomena like wellbeing, it is crucial to recognize the nuanced distinction between measuring the overall phenomenon of wellbeing and its specific contributing factors (Blijlevens et al., 2017). This means separating general assessments of wellbeing from the various antecedents, local indicators, and elements that shape it. For instance, while the Satisfaction with Life Scale measures overall cognitive judgments of life satisfaction, the Scale of Positive and Negative Experiences (SPANE) specifically targets affective feelings like joy, sadness, anger, and stress (Diener et al., 2010). Together, these scales provide a richer, multidimensional assessment by measuring distinct components that constitute subjective wellbeing. However, researchers must be careful not to conflate measures of specific feelings or experiences with holistic wellbeing itself. Wellbeing arises from a complex interplay of many factors, so measures must clearly distinguish between the overall state versus its multiple drivers. Keeping this distinction clear by using validated scales for both global wellbeing and domain-specific elements enables deeper investigation of how these components interact.

Broadly speaking, wellbeing assessment can be categorized into objective and subjective approaches. Objective measures include income, housing quality, employment status, and physical health indicators. These aim to assess wellbeing from external criteria (Voukelatou et al., 2021). In contrast, subjective measures currently rely on self-reports of individuals' own evaluations of their wellbeing. These include assessing life satisfaction, positive and negative affect, purpose in life, harmony in life, and other psychological dimensions (Cooke, Melchert, & Connor, 2016; Linton, Dieppe, & Medina-Lara, 2016).

The assessment of wellbeing has expanded significantly in recent decades across research, clinical, and policy contexts (Linton et al., 2016). However, wellbeing remains a complex, multidimensional concept lacking consensus regarding definition and measurement (Cooke et al., 2016). Therefore, various theoretical approaches have been applied, including hedonic, eudaimonic, quality of life, and wellness models. This has resulted in a proliferation of assessment instruments, with Cooke et al. (2016) identifying 42 instruments to assess wellbeing or related constructs and Linton et al. (2016) identifying 99 instruments designed to assess self-reported wellbeing.

These instruments vary widely in their conceptualization and operational-

ization of wellbeing. For instance, hedonic measures tend to focus on life satisfaction and positive/negative affect, with examples including the Satisfaction with Life Scale (Diener et al., 1985), Subjective Happiness Scale (Lyubomirsky & Lepper, 1999), and Happiness Measures (Fordyce, 1988). In contrast, eudaimonic measures emphasize fulfillment of potential and functioning, such as the Questionnaire for Eudaimonic wellbeing (Waterman et al., 2010) Ryff's Scales of Psychological wellbeing (Ryff, 1989), the PERMA-profiler (Butler & Kern, 2016). In addition to these two categories, Cooke et al. (2016) identify two more. Namely, Quality of life measures take a broader biopsychosocial perspective, including the WHO Quality of Life Scale (Group, 1998) and Comprehensive Quality of Life Scale (Cummins, Mccabe, Romeo, & Gullone, 1994). Finally, wellness measures incorporate factors like nutrition, fitness, and spirituality, such as the Wellness Inventory (Abrahams & Balkin, 2006) and Optimal Living Profile (Renger et al., 2000).

In conclusion, assessing wellbeing in relation to AI requires nuanced models capturing the complex interplay of factors influencing human flourishing. This involves developing contextualized understandings of wellbeing that consider how positive and negative experiences shape us. It also requires achieving a balanced perspective on individual and collective wellbeing given cultural variation in self-construals. Most crucially, adopting culturally sensitive, participatory methods is critical to move beyond decontextualized definitions toward localized constructs of wellbeing. In this way, an interdisciplinary approach integrating qualitative and quantitative data can elucidate the pathways to human flourishing relevant when evaluating AI's impacts.

2.2.8. Why should wellbeing be an objective for AI?

Now that we have an overview of what wellbeing is, there are four main reasons why it should be a key objective in AI development:

1. **Increasing ubiquity**: The integration of AI into our daily lives is rapidly increasing, granting these systems a significant influence over various aspects of human wellbeing. According to an IEEE standards review, this integration necessitates that AI not only aligns with but actively enhances human wellbeing, as its decisions and actions can profoundly permeate many facets of our lives (Shahriari & Shahriari, 2017).

2. **Emerging evidence**: There is emerging evidence, albeit not yet

conclusive, suggesting that wellbeing might be impacted—both positively and negatively—by AI systems (Havrda & Klocek, 2023). This highlights the importance of designing AI to safeguard and promote wellbeing to prevent potential harm.

3. **Sociotechnical responsibility**: AI systems represent unique sociotechnical systems possessing agency (van de Poel, 2020); as such, they hold a special responsibility to respect and uphold human dignity and rights. AI development must prioritize these aspects, ensuring that AI systems contribute to, rather than detract from, the societal and ethical values that underpin human dignity (Floridi et al., 2018).

4. **Tremendous potential**: AI holds tremendous potential for positive impact, akin to major historical revolutions like the agricultural or industrial revolutions (Chui et al., 2023; K. Crawford, 2021; Gates, 2023; Schwab, 2017; Shahriari & Shahriari, 2017; Shneiderman, 2022). Harnessing this potential responsibly can lead to substantial advancements in human wellbeing. This underscores the necessity of directing AI development towards outcomes that not only avoid harm but also actively contribute to the betterment of society.

It is clear why wellbeing *should* be an objective, but the question remains whether it *could* be one that AI systems can reasonably optimize for.

2.2.9. Are metrics of wellbeing an appropriate objective for AI?

The established history of wellbeing measurement supports the suitability of aligning AI with wellbeing objectives. According to Stray (2020), this historical basis makes wellbeing a robust and meaningful target for AI alignment. Further emphasizing this trend, the 2020 IEEE standard, discussed by Schiff, Ayesh, Musikanski, and Havens (2020), advocates for standardized wellbeing assessments in AI applications. This standard marks a paradigm shift in AI design, suggesting a more holistic approach to technology development (Schiff, Murahwi, Musikanski, & Havens, 2019). In a more focused perspective, Musikanski et al. (2020) advocate for prioritizing community wellbeing within AI research. They argue that by integrating community wellbeing considerations, AI development can address the multifaceted challenges and opportunities it presents to various communities more effectively.

But why should wellbeing be the focal point rather than other human values? S. Harris (2010) contends that wellbeing is the only morally

defensible objective for alignment. This argument is predicated on the understanding that any system interacting with humans—whether AI-centered systems like social media, streaming platforms, and chatbots, or traditional institutions like schools, hospitals, and governmental entities—ought to be in harmony with human values. However, the pursuit of aligning with human values often entails complex tradeoffs (Gabriel, 2020; Stray et al., 2023). These tradeoffs manifest in the tension between corporate and ethical motives and among ethical principles. For example, measures that enhance our sense of safety, such as facial recognition technology in public spaces, may concurrently infringe upon our autonomy or privacy. This complexity raises critical questions about which values should be prioritized and the basis for such prioritization.

In this context, S. Harris (2010) puts forth wellbeing as an overarching guiding principle. He argues that wellbeing depends on and encompasses all other values. To determine what to prioritize, he proposes the notion of a "moral landscape" to navigate. This landscape comprises peaks of human flourishing and valleys of suffering. Harris argues we should chart a course focused on reaching the highest peaks. And to inform which route leads to the peaks, we must measure wellbeing to predict the outcomes of various trajectories. Therefore, when facing prioritization questions, we should emphasize those specific values and value tradeoffs that empirically further wellbeing.

Some critics contend that the concept of wellbeing is too indistinct or subjective to reliably ground moral frameworks. In response, S. Harris (2010) employs an analogy with the field of medicine to address this skepticism. He identifies three potential challenges to using wellbeing as a foundation for morality, which could similarly be applied to the concept of 'health' in medicine.

- **Value Problem**: Just as improving health is valued in medicine, improving wellbeing should be valued in morality. There's no need to scientifically justify why health or wellbeing are important—they are self-evidently valued;
- **Persuasion Problem**: Some people don't care about health or wellbeing, but that doesn't undermine medicine or morality as fields. We still have medical truths even if some reject them. Similarly, moral truths exist even if some reject caring about wellbeing;
- **Measurement Problem**: Both health and wellbeing are difficult to precisely define and measure, but that doesn't prevent medicine from

> being practiced scientifically. Similarly, a loose definition of wellbeing doesn't preclude a scientific understanding of morality;

In essence, S. Harris (2010) argues that many of the criticisms against basing morality—or moral questions such as what constitutes ethical AI—on wellbeing could also be made against medicine's basis in health. But we don't doubt medicine as a science simply because health is hard to measure or because some people don't care about it. So we shouldn't question the prospect of a science of morality solely for those reasons regarding wellbeing. Accepting this position, the question then become what may be a good metric for wellbeing. Could we operationalize wellbeing along a single axis or take a behavioral metric, feed it into an AI system, and expect it to promote wellbeing?

The answer is no; and this is where 'Goodhart's Law' becomes particularly relevant. Goodhart's Law prescribes that when a measure becomes a target, it ceases to be a good measure (Goodhart, 1975). This principle highlights the inherent risk of reducing wellbeing to a mere metric for AI optimization. When focused on optimizing specific metrics, it may lead to unintended consequences, such as manipulation, gaming, and a narrow focus on short-term goals. This can result in adverse outcomes that diverge from the intended goals of the AI application (Thomas & Uminsky, 2020). Additionally, adopting a singular theoretical paradigm of wellbeing risks excluding dimensions not emphasized by that specific framework. A multidimensional perspective is needed to capture the breadth of factors influencing human flourishing. Comprehensively understanding the complexity of human experience is a non-trivial task. Still, human-centered design can greatly contribute to it by emphasizing integrative practice and attuning to people's lived experiences. This leads to our next section, addressing the role of HCD. Before that, a brief intermezzo addressing how AI may affect wellbeing.

2.3. Intermezzo: How might AI affect wellbeing?

Now that we understand what AI and wellbeing are, how may they interact? Some examples exist, such as in the Social Dilemma (Orlowski, 2020). However, the current literature lacks consensus regarding the general effects of such platforms. This may stem from limitations in existing wellbeing metrics or the scope of analyses conducted. To provide a concise overview of how this research envisions potential effects within its framed

context, the table below presents a basic mapping of aspects of wellbeing as conceptualized by (Ryff, 1989)'s model of Psychological Wellbeing (PWB) to affordances of Netflix. While superficial, this mapping aims to concretize discourse around AI and wellbeing through relatable examples.

Table 2.2: Examples of the potential effect of Netflix on wellbeing using Ryff (1989)'s model of PWB

Dimensions of wellbeing	Positive Effects	Negative Effects
Self-Acceptance	Offers a diverse range of content promoting inclusivity and representation, enhancing viewers' understanding and acceptance of different cultures and identities, and promoting self-awareness and acceptance.	Binge-watching can lead to feelings of guilt and self-criticism, especially when it comes at the expense of other responsibilities or self-care activities.
Positive Relations	Facilitates shared viewing experiences, either physically or virtually, helping to strengthen bonds through shared interests and discussions about content.	Excessive viewing can lead to social isolation, reducing time spent on building and maintaining real-life relationships.
Autonomy	Empowers viewers with the choice of what, when, and how much to watch, fostering a sense of control over their entertainment choices, and encourages independent decision-making in entertainment choices.	The algorithm-driven suggestions can create a dependency on the platform for entertainment choices, potentially limiting viewers' exposure to diverse content.
Environmental Mastery	Provides educational content and documentaries that can enhance viewers' understanding of the world and empower them with knowledge to navigate various life situations.	Prolonged periods of inactivity associated with binge-watching can lead to a neglect of personal environment management and physical health.
Purpose in Life	Inspiring stories and documentaries can motivate viewers to pursue their goals and passions, providing direction and a sense of purpose, and highlight various life paths and careers.	Excessive consumption can distract from personal goals and ambitions, leading to procrastination and a lack of focus on personal development.
Personal Growth	A broad range of genres and topics can contribute to viewers' cultural and intellectual development, enhancing their understanding of the world and different perspectives, and enabling users to expand their knowledge and cultural understanding.	Prolonged screen time and passive consumption can limit personal growth opportunities that come from active engagement and real-world experiences.

2.4. The role of human-centered design

Having established an understanding of AI and how wellbeing can act as its alignment objective, the next step is to explore the role of designers in navigating this complex landscape. Firstly, we'll discuss the role to play in the former—AI design—then, more specifically with respect to wellbeing.

In his 2020 exploration of human-centered design's (HCD) integration in AI development, Auernhammer (2022) underscores the significance of addressing ethical, accountability, and practicality challenges often missed by purely engineering-focused methods. He advocates for a comprehensive, interdisciplinary approach that merges technological expertise with social sciences and humanities insights. Such a blend is intrinsic to HCD, positioning it as a pivotal contributor in various crucial capacities:

1. **Examining societal/ethical implications:** approaches like human-centered systems and social design can reveal impacts of AI on social systems and ethical dynamics;
2. **Representing diverse perspectives:** participatory design and inclusive design can incorporate input from different stakeholders and communities;
3. **Understanding human needs and behavior:** Interaction design, persuasive technology, and need-design response approaches focus on human needs, emotions, motivations, and responses that should inform AI design;
4. **Prototyping and testing:** Interaction design supports assessment of AI systems through prototyping and user testing to guide iterative refinement;
5. **Envisioning beneficial applications:** Framings like human-centered computing encourage focusing innovation on enhancing human capabilities and wellbeing.

While vital roles exist for human-centered design in developing ethical, socially-responsible AI, integrating HCD principles into AI systems faces challenges. As Yang, Steinfeld, Rosé, and Zimmerman (2020) point out, factors like uncertainty of capabilities and complexity of outputs create barriers to traditional user-centered processes reliant on clear requirements and predictable behaviors. This makes practices like representing diverse perspectives, prototyping, and testing difficult when working with constantly-adapting AI systems. Additionally, the mismatch between AI's

emergent learning capacities and HCD's more fixed notions of design goals hinders approaches for envisioning beneficial applications (Sadek, Calvo, & Mougenot, 2023a). While HCD is well-equipped to examine societal implications and human needs related to AI, reconciling open-ended AI capacities with human-centered values requires a willingness to re-examine some conventional assumptions. Rather than rejecting AI's uncertainties, HCD must embrace its emergent qualities through processes allowing for deep collaboration at every stage. This entails fundamentally reassessing rigid interpretations of needs analysis, prototyping fidelity, and design requirements in an AI context.

Despite these challenges, an emerging body of work explores integrating human-centered design into AI development, underscoring designers' crucial role. For example, Yildirim's research (Yildirim et al., 2022; Yildirim, Oh, et al., 2023; Yildirim, Pushkarna, Goyal, Wattenberg, & Viégas, 2023) examines how designers collaborate on enterprise AI teams—fostering partnerships with data scientists, engaging stakeholders, developing creative tools, and applying guidelines impactfully. Other works emphasize direct stakeholder participation through methods like co-design (Zhang, Boltz, Lynn, Wang, & Lee, 2023), interactive probes, and service blueprinting (Li & Lu, 2021) to center user perspectives and contexts. Additionally, reviews like Sadek et al. (2023a) synthesize learnings on conversational agent co-design into best practices. Collectively, this research reveals pathways for designers to steer AI innovation responsibly and beneficially through collaboration, participation, and translating human values into implementations.

Further, the literature contains limited (mature) work examining the intersection of design, AI, and wellbeing. As discussed previously, Stray (2020)'s analysis of using wellbeing metrics to optimize social media platforms suggests promise but also open questions regarding implementation. Additionally, Calvo, Peters, Vold, and Ryan (2020) demonstrated the need to respect user autonomy in AI systems like YouTube's recommender. However, cases that actually design and assess wellbeing-focused interventions remain sparse. Some conduct this work under the guise of Digital Wellbeing, however this is not the same as Positive AI. For example, reviewing the digital wellbeing track of 2023's CHI conference shows most articles focus narrowly on dark patterns, intentionally addressing only one system aspect and concentrating solely on harm mitigation rather than active enhancement of wellbeing through positive design (e.g., Chordia et al., 2023; Mildner, Savino, Doyle, Cowan, & Malaka, 2023; Monge Roffarello, Lukoff, & De Russis, 2023). The point is that, in the broader field of digital wellbeing, there's a notable emphasis on mitigating harm from digital spaces rather than

actively using them to enhance wellbeing (e.g., Cecchinato et al., 2019; Monge Roffarello & De Russis, 2019; Vanden Abeele, 2021).

More closely related work in this space includes Lukoff et al. (2023), who designed a proof-of-concept system to foster agency, and Lyngs et al. (2020), who assessed interventions for control in Facebook. While promising, these could benefit from a broader perspective, engaging communities to conceptualize context-specific notions of wellbeing. That is, the question of whether agency, for example, is essential in other contexts is unaddressed—which is what the field needs: "translational work" (R. Wong, Madaio, & Merrill, 2022). In contrast, the hospitality industry study by (Spektor et al., 2023) provides a tangible example of how worker feedback can directly inform the adaptation of AI systems, suggesting a more holistic approach to wellbeing in AI design. This provides a perfect segue into what I envision as Positive AI, building on existing 'Positive' traditions.

2.4.1. Positive Design, Technology & Computing

More specifically, Positive Design has emerged as a transformative approach that pivots the focus from merely solving problems to enhancing the subjective wellbeing of individuals (Desmet & Pohlmeyer, 2013). At its core, Positive Design is guided by the explicit aim of fostering a lasting appreciation for life, intertwining three integral components: creating pleasure, nurturing personal significance, and encouraging virtuous behavior. This philosophy goes beyond the traditional problem-focused design paradigm by embracing a possibility-oriented approach. Here, the emphasis is on augmenting existing potential and creating new opportunities for growth and development.

Complementing this, Positive Technology (Gaggioli, Riva, Peters, & Calvo, 2017), Positive Computing (Calvo & Peters, 2014), and Design for Wellbeing (Calvo & Peters, 2019) concentrate on integrating wellbeing research into technology to promote human flourishing. Beyond mitigating harm, these interrelated fields envision creating ethical technologies that respect user needs and encourage thriving. A key outcome is the Motivation, Engagement, and Thriving in User Experience (METUX) model, which draws on Self-determination Theory (SDT) (Deci & Ryan, 2008b) to reveal how technology can support or undermine basic psychological needs, shaping user motivation, engagement, and wellbeing (Calvo & Peters, 2014). This model was applied by Calvo, Peters, and Cave (2020) to scrutinize human-AI interactions across six spheres: Adoption (analyzing how users decide to use AI), Interface (assessing user interaction with AI systems), Task (evaluating specific tasks facilitated by AI and their impact on psychological

needs), Behaviour (considering broader behaviors promoted by AI and their alignment with personal goals), Life (exploring the influence of AI on overall life and wellbeing), and Society (examining broader societal implications of AI, including ethical concerns). This detailed application of METUX enables a comprehensive understanding of how each aspect of AI technology use supports or impairs psychological needs, ultimately influencing user motivation, engagement, and overall wellbeing (Calvo, Peters, Vold, & Ryan, 2020).

These fields lay the foundation for what this dissertation refers to as "Positive AI," which echoes the emphasis on leveraging technology to promote human flourishing through a focus on wellbeing outcomes. This work focuses on what Calvo and Peters (2014) term "active integration" (see Table 2.3), concentrating on commercially-operated AI systems like major platforms where business objectives are the primary goal. The motivation is that such widely used systems have tremendous potential for impact, both positive and negative. Thus, inspiring their ethical realignment is imperative. However, their commercial nature poses distinct challenges compared to dedicated wellbeing systems since financial incentives will compete with social ones. Additionally, issues in dedicated systems compound in active ones. Therefore, uncovering techniques for the latter can inform broader integration strategies. In essence, while harder, transforming commercial AI to prioritize societal wellbeing over profits promises immense benefits to human welfare.

Table 2.3: Positive Computing Strategies adapted from Calvo and Peters (2014)

Positive computing strategies	
Not positive design	Wellbeing and human potential were not considered in the design of the technology
Preventative integration	Obstacles or compromises to wellbeing are treated as errors
Active integration	A technology that is designed to actively support components of wellbeing or human potential in an application that has a different overall goal
Dedicated integration	A technology that is purposefully built to and dedicated to fostering wellbeing and human potential in some way

In summary, this background section has outlined the core concepts grounding this dissertation—artificial intelligence, wellbeing, and human-centered design—while situating it within the context of Positive AI. It has examined ethical and technical efforts to align AI with human values, explored multifaceted notions of wellbeing, and discussed pathways for human-centered design to ensure innovation remains responsive to human experience. Building on these foundations, the subsequent chapters present original research contributing to the advancement of AI focused on actively enhancing human flourishing through an assessment-driven, participatory approach centered on wellbeing.

3

Seven key challenges

Despite aspirations for AI to promote human wellbeing, significant obstacles still impede this vision from becoming reality. Building on the cybernetic perspective proposed in Section 2.1.1, this chapter draws on that conceptual grounding to organize current challenges in designing AI to support wellbeing. Using this lens, it structures key challenges around processes of modeling, assessing, designing for, and optimizing wellbeing. Specifically, it outlines barriers in 1) identifying relevant wellbeing dimensions and 2) attributing causal relationships; 3) operationalizing wellbeing and 4) translating qualitative insights into quantitative system metrics; 5) designing effective interventions across sociotechnical system levels; 6) managing optimization tradeoffs; and 7) conflicting pace layers. By delineating these knowledge gaps, the chapter provides a research agenda for advancing scientific understanding and practical implementation of Positive AI. It argues for centralizing participatory assessment to enable responsive alignment as contexts evolve. The discussion informs a subsequent case study (Chapter 4) that directly responds to challenges through an iterative, human-centered approach emphasizing stakeholder engagement. In this way, the chapter bridges conceptual foundations to actionable directions for realizing AI's potential to foster flourishing. Because of the paper-based structure of this dissertation, the beginning of this chapter discusses themes (e.g., AI alignment, Ethical AI, the role of HCD) extensively addressed in the Introduction.

This chapter is under review for publication at *She Ji: The Journal of Design, Economics and Innovation* as an article titled "Positive AI: Key Challenges for Designing Artificial Intelligence for Wellbeing"

Artificial Intelligence (AI) is a double-edged sword: on one hand, AI promises to provide great advances that could benefit humanity, but on the other hand, AI poses substantial (even existential) risks. With advancements happening daily, many people are increasingly worried about AI's impact on their lives. To ensure AI progresses beneficially, some researchers have proposed "wellbeing" as a key objective to govern AI. This article addresses key challenges in designing AI for wellbeing. We group these challenges into issues of modeling wellbeing in context, assessing wellbeing in context, designing interventions to improve wellbeing, and maintaining AI alignment with wellbeing over time. The identification of these challenges provides a scope for efforts to help ensure that AI developments are aligned with human wellbeing.

3.1. Introduction

The rapid advancement and adoption of generative AI (GenAI) technologies like *ChatGPT* signify the dawn of "The Age of AI." (Gates, 2023; Kissinger et al., 2021) These developments mark a significant leap in the capabilities and adoption of AI systems. However, for many people, the swift and disorienting integration of AI into daily life raises many issues (Cugurullo & Acheampong, 2023; Fietta, Zecchinato, Stasi, Polato, & Monaro, 2022; Qasem, 2023). Concerns include the potential impacts on employment, privacy, and inequality, along with broader societal implications like human rights, mental health, and the preservation of democratic norms (Future of Life Institute, 2023; Prabhakaran, Mitchell, Gebru, & Gabriel, 2022; Shahriari & Shahriari, 2017; Stray, 2020). This article argues for the importance of wellbeing as a key objective in AI and for human-centered design (HCD) as a key methodology. Based on this framing, it shares a set of key challenges that will face designers of AI for wellbeing, or *Positive AI*.

The idea that AI should support wellbeing is not uncommon. In 2018, Zuckerberg (2018) (CEO of Meta, previously Facebook) publicly stated that wellbeing should be the goal of AI. Further, in an interview Jan Leike (Wiblin, n.d.) (head of the 'Superalignment' research lab at OpenAI) said AI optimization should align to "flourishing." Wellbeing, however, is complicated. It is not a naturally observable quantity, but rather a multifaceted construct that is based, at least in part, on conscious human experiences (Ruggeri, Garcia-Garzon, Maguire, Matz, & Huppert, 2020). Therefore, designing *Positive AI* requires understanding and shaping human experiences. This situates the challenge squarely in the domain of human-centered design (Auernhammer, 2022). Before reviewing the possibilities

for HCD in designing AI for wellbeing, we will briefly address other fields associated with the creation of positive human outcomes in AI. The current article is not the venue for reviewing them in-depth. Yet, we find it important that Positive AI designers are broadly aware of their contributions.

3.1.1. Ethical AI

AI ethicists have been formulating ethical principles and frameworks to responsibly guide the development and implementation of AI systems. A key contribution came from Floridi et al. (2018) who synthesized a set of core ethical principles like beneficence, non-maleficence, autonomy, justice, and explicability for cultivating a "Good AI Society." However, as principles alone they are insufficient (Mittelstadt, 2019): they need to be translated into concrete practices. Much work remains to develop these into practical tools and methodologies (Morley et al., 2020). Recent work has begun exploring approaches for embedding ethical values directly into AI system design (Klenk & Duijf, 2021; van de Poel, 2020), from which Value-sensitive Design (VSD) has emerged as a candidate to bridge the principle-practice gap (Umbrello & van de Poel, 2021). A recent critical review (Sadek et al., 2023b) indicates that VSD may be effective but limited. Some limitations include the inadequate elicitation of values, a tendency to depend on pre-established values over context-specific ones, and a lack of precise instructions for embedding values.

3.1.2. AI Alignment

Considering the potential for harm done by AI, some refer to such systems as misaligned with human values. For example, referencing social media, ethicist Tristan Harris says that by optimizing for attention, these platforms are misaligned with human wellbeing and dignity (T. Harris, 2017). "AI alignment" is a field of research that aims to develop systems that are aligned with human values and intent (Christian, 2020). Alignment has been earlier studied as the principal-agent problem in economics and law, where an agent must achieve the objectives and interests of the principal (Hadfield-Menell & Hadfield, 2019). For example, in a car repair scenario, the car owner (principal) expects the mechanic (agent) to fix the car efficiently and affordably. However, the mechanic might suggest unnecessary repairs to increase the bill (misalignment), contrary to the owner's desire for cost-effective service. Taking this framework to AI systems, alignment means ensuring that AI agents effectively and reliably pursue the goals and preferences set by their designers and users. One successful example

of technical alignment work is the use of Reinforcement Learning with Human Feedback (RLHF) (Christiano et al., 2017), which uses human preference data to align the behavior of Large Language Models (LLMs). Related techniques include Constitutional AI (Bai et al., 2022) and inverse reinforcement learning (IRL) (A. Y. Ng & Russell, 2000). There are many new techniques in the expanding field of technical AI alignment.[1] However, while these technical efforts show tremendous progress, their technology-centered perspective risks missing broader sociotechnical considerations, such as the design of human systems to effectively respond to AI (Dung, 2023).

3.1.3. Why human-centered design?

Given the intrinsic relationship between wellbeing and conscious experience, some scholars have argued for the importance of human-centered design (HCD) in AI (Calvo & Peters, 2014; Desmet & Pohlmeyer, 2013). One reason is that, as a field, HCD focuses on understanding and shaping human experiences. However, there are a variety of ways in which HCD might complement ethical perspectives and address gaps in the AI alignment field. For instance, HCD might help bring concrete implementation methods and a broader systemic perspective. A core tenet of HCD is to prioritize the needs, values, and capabilities of users, ensuring that the design process is centered around human beings and their interactions with technology.

Designers are trained to attend to—and empathize with—human experiences (Norman, 2013). This means considering the full context surrounding users and technologies, rather than just narrow functionality, as well as prioritizing the understanding of diverse users' needs and experiences from their point of view (Sanders & Stappers, 2008). They are equipped with the ability to engage in stakeholder participation and reveal the ethical priorities and deeply-held beliefs relevant to design projects (Zhang et al., 2023). This encompasses a blend of competencies from engineering design, including problem definition, scoping, and rapid prototyping, combined with methodologies from social sciences like conducting ethnographic research, interviews, deriving understanding from qualitative data, and engaging in empathetic practices (Kramer, Agogino, & Roschuni, 2016).

These skills are particularly important for AI because the integration of diverse perspectives ensures that both technical efficiency and societal impacts are considered in AI development (Auernhammer, 2022). For instance, experimentation and prototyping in AI benefit from this blend,

[1] For a comprehensive overview, readers are referred to the live agenda summarizing ongoing alignment efforts posted to the AI Alignment Forum (Technicalities & Stag, 2023), which has been founded by prominent alignment researcher Eliezer Yudkowsky.

allowing for iterative refinement and alignment with human needs and values. Prototyping in AI can be difficult because of the inherent unpredictability and complexity in AI's capabilities and outputs (Yang et al., 2020). HCD may help by applying user-focused approaches to manage these uncertainties. Moreover, involving end users directly in the design process ensures that AI solutions are tailored to real-world requirements, making the technology more accessible, usable, and effective (Li & Lu, 2021; J. Zhu, Liapis, Risi, Bidarra, & Youngblood, 2018). In summary, the HCD perspective can complement existing ethical and technical viewpoints in AI development, as it offers methodologies to create systems that balance technical robustness with socially responsible outcomes that benefit people and society at large.

The field of Positive Design focuses on promoting human flourishing. The Positive Design Framework provides a scaffold for solutions that can enhance subjective wellbeing through components like pleasure, meaning, and virtue (Desmet & Pohlmeyer, 2013). Grounding positive design in theory and evaluating its effect through controlled studies helps ensure that designed solutions truly contribute to people's happiness. Similarly, the Positive Computing (Calvo & Peters, 2014, 2019; Gaggioli et al., 2017) movement aims to leverage technology to measurably improve wellbeing and human potential. The emphasis on collaborations between fields like psychology, computer science, and design in positive computing underscores the importance of an interdisciplinary, human-centric approach for developing AI focused on wellbeing objectives (Calvo, Vella-Brodrick, Desmet, & M. Ryan, 2016). In many ways, the tenets of positive design and positive computing have helped lay the foundation for what we now call "Positive AI."

3.1.4. Why wellbeing?

A growing movement of scholars advocates for the incorporation of wellbeing metrics into AI systems so that optimization efforts can measurably contribute to social benefit (Schiff, Ayesh, et al., 2020; Shahriari & Shahriari, 2017). Specifically, they argue that measures of wellbeing can help manage AI's effects on society (Musikanski et al., 2020). Indeed, wellbeing has a strong methodological foundation (Stray, 2020), and there is extensive research on defining and measuring wellbeing; this suggests that algorithmic systems may be able to systematically optimize wellbeing (Havrda & Rakova, 2020).

Wellbeing's complexity captures many relevant societal concerns AI systems should address (Stray, 2020). This combination of rich meaning and inherent measurability supports the operationalizing wellbeing as an optimization objective for AI systems. This sentiment is also expressed

by a recent IEEE standards review that argues for the adoption of holistic wellbeing frameworks (like IEEE 7010) to guide the design, deployment, and evaluation of AI systems (Schiff, Ayesh, et al., 2020). However, significant questions remain regarding whether available wellbeing frameworks are fully sufficient, whether existing metrics are sufficient, what the impacts of wellbeing optimization may be (Musikanski et al., 2020; Schiff, Ayesh, et al., 2020; Stray, 2020).

Some argue that wellbeing is a sort of ultimate objective: in *The Moral Landscape*, S. Harris (2010) argues that other values like fairness, transparency, or accountability should be seen as components that contribute to wellbeing, rather than ends in themselves. From this perspective, optimizing for wellbeing involves optimizing for all values that matter, but only insofar as they contribute empirically to wellbeing. In so far as AI systems are able to assess their own impact on human wellbeing, they may be able to potentially maximize all benefits and minimize all harms experienced by users and society (Havrda & Klocek, 2023). Wellbeing optimization might then allow for the management of complex issues like misinformation and inequality associated with AI systems (Stray, 2020).

3.1.5. Framing the challenges: human-centered design of AI systems

As a term, 'Artificial intelligence' is used to describe both a characteristic of computer systems and the methods employed to develop this feature, such as machine learning (ML) (Gabriel, 2020). Intelligence in both humans and machines has been defined as "an agent's general ability to achieve goals in a wide range of environments." (p. 9 Legg & Hutter, 2007) Following this definition, AI researchers Stuart Russel and Peter Norvig define *artificial* intelligence as a *designed* agent that perceives its environment through sensors and acts upon that environment using actuators (Russell & Norvig, 2022). The result of these sensors and actuators is a feedback loop that incorporates system output (e.g., action in its environment) as input for its future actions (e.g., the action had the desired effect). A cybernetic perspective examines these broader feedback loops between AI systems, their environment, and the social context in which they operate. Thus, AI *systems* (in contrast to AI/ML algorithms) can be viewed as sociotechnical systems embedded within a complex network of feedback loops (van de Poel, 2020). This broader and more systemic view of AI has been proposed as an approach to deal with some of the challenges of current and future AI systems (Dobbe et al., 2021; Krippendorff, 2021; van der Maden et al., 2022).

A human-centered design perspective enables designers to look beyond

"the algorithm" to consider how AI interacts within a network of social, ethical, cultural, and political factors (van de Poel, 2020). This means considering how AI influences human behavior, societal norms, and institutional structures—and how AI is, in turn, influenced. This perspective requires engaging with diverse stakeholders to understand their values and needs iterative and reflective design processes that continually assess and respond to these complex dynamics (Sadek et al., 2023b).

This perspective creates new affordances for the design of AI for wellbeing. Shaping the impact of AI can occur through multiple components of the AI system, including the technical artifacts, institutions, practices, etcetera—all in addition to the design of algorithms. For example, consider the role of AI in a video streaming platform like Netflix. A focus on the algorithm is limiting because one can only affect the likelihood of a particular recommendation. In contrast, a broader perspective opens up different kinds of interventions. For instance, in the user interface (e.g., autoplay); the organizational level (e.g., establishing boardroom content acquisition metrics beyond just engagement and growth); data science (e.g., introducing new metrics for optimization that prioritize suggesting content from more diverse voices); or the broader ecosystem (e.g., funding initiatives to broaden representation in the creative industry talent pipeline). In other words, design interventions can occur at multiple levels of the AI *system*, not just in the algorithm. Designers can even consider interventions outside of the control of the AI platform, such as a 'Netflix Watch Club' or an alternative YouTube user interface for education (Lukoff et al., 2023).

Why is this broadened view important? Rather than aligning AI to the needs of society, there may be many cases where it may be more appropriate for social institutions to adapt to AI. For instance, while ChatGPT could potentially be aligned with the needs of K12 schools (i.e. so that students are prevented from cheating on their assignments), designers may wish to create new guidelines for positive integration of AI in their courses (e.g., promoting AI literacy in using ChatGPT (Mollick, 2023)). With this systemic perspective, there are expanded opportunities for guiding AI impacts beyond the algorithmic design itself. This shows how there are new opportunities for creating AI for Wellbeing beyond what is typically the scope of AI alignment or ethical AI research.

In this article, we conceptualize AI as a sociotechnical system involving a complex interaction of various feedback loops, each optimized for specific objectives within the system. The essence of 'Positive AI' lies in harmonizing these objectives towards a singular, overarching aim: enhancing wellbeing.

3.2. Key Challenges

As experienced design researchers in this domain, we have consistently encountered unique challenges in designing *Positive AI*. The challenges we outline here are intended to inform and guide other designers embarking on similar ventures. This conceptual framework aids in structuring the challenges of designing AI for wellbeing around the following questions:

1. **Modeling** the state of the system: How do we operationally define wellbeing within the context of a particular sociotechnical system?
    - For example: *What wellbeing dimensions are important in the context of Netflix and how do we attribute changes in wellbeing to components of the system?*
2. **Assessing** the state of the system: How do we translate qualitative experiences into assessment metrics?
    - For example: *How can we elicit how people feel about their interactions with TikTok, and how can these experiences be translated into metrics that can be used for assessing future interventions and optimization processes?*
3. **Designing** system actuators: How do we design interventions in AI systems that promote and enhance wellbeing?
    - For example: *How do we know where in the sociotechnical system of ChatGPT we should and can intervene, and how do we know whether our potential interventions will achieve the desired effect?*
4. **Optimizing** the system objective: How do we know whether we are getting close to our desired goal?
    - For example: *How do we manage tradeoffs between autonomy and social connection in designing for wellbeing on Reddit, and how do align immediate outcomes with long-term wellbeing goals?*

We have found the concept of cybernetics to be a useful lens for organizing the challenges of designing AI for wellbeing. Cybernetics provides a systemic and holistic viewpoint that defines clear mechanisms for impact (Tabari, 2022). For us, a cybernetic perspective naturally accommodates ecological and sociotechnical perspectives on the numerous feedback loops governing human and AI systems in society today. This aligns with suggestions from other scholars who have proposed adopting a cybernetic approach to mitigate the negative effects of AI and deal with the complexities of

the challenges surrounding AI (Dobbe et al., 2021; Krippendorff, 2021; Sato, 1991). They argue that a cybernetic perspective, as opposed to a techno-centered one, allows for the inclusion of ecological perspectives. This cybernetic perspective allows us to organize the key challenges of designing AI for wellbeing into four main categories, as shown in Table 3.1. This table serves as an overview and guides the structure for the rest of this chapter, with the numbers corresponding to the categories depicted in Figure 3.1.

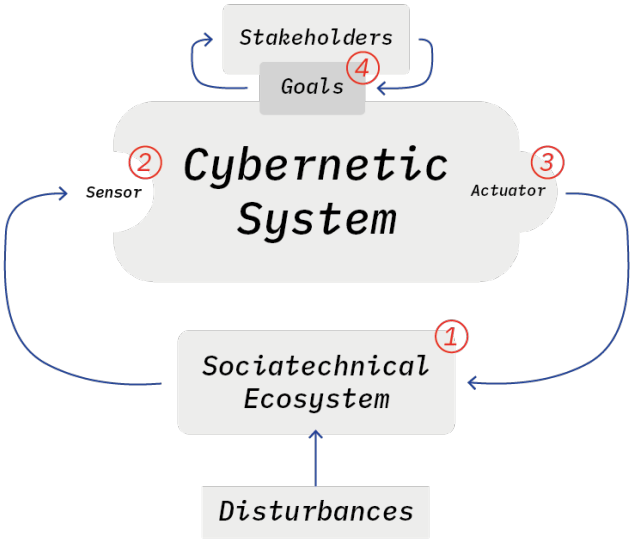


Figure 3.1: Shows a schematic representation of a cybernetic system. The different challenges can be mapped onto this framework: (1) understanding the system context which entails modeling the relation between wellbeing of the systems constituents and its various components; (2) operationalizing said model of wellbeing; (3) designing interventions to actively promote operationalized model of wellbeing; and (4) retaining alignment with the overall goal. The latter refers to both challenges of algorithmic optimization as well as scrutinizing the objective (e.g., is the wellbeing objective still aligned to needs and desires of all relevant stakeholders?)

Table 3.1: Table 2. Overview of Key Challenges

Category	Challenge	Explanation
How to model wellbeing?	(1a) Identifying relevant wellbeing dimensions	There is a long tradition of wellbeing research spawning many different theoretical paradigms. In any given context, the appropriate theoretical paradigm may vary.
	(1b) Modeling wellbeing and interpreting fluctuations	The multifaceted nature of wellbeing and the richness of the world in which we live, make it difficult to model fluctuations in wellbeing to (specific features of) a system.
How to assess wellbeing?	(2a) Operationalizing wellbeing	In order for a system to respond properly and timely, we have to measure wellbeing in a dynamic, continuous manner that is sensitive to the context.
	(2b) Translating qualitative experiences into system metrics	We currently lack methods for connecting small-scale qualitative research on wellbeing to large-scale, quantitative methods, making it difficult to ensure that metrics of wellbeing are aligned with human experiences.
How to design for wellbeing?	(3) Designing interventions to promote wellbeing	It is not straightforward how an algorithm, the UI of a platform, or its content may lead to wellbeing effects. There's both a lack of examples and a lack of appropriate design methodologies for Positive AI.
How to optimize for wellbeing?	(4a) Optimization tradeoffs and prioritization	Optimization causes tradeoffs that are difficult to measure, making it difficult to balance different goals and predict the effects of novel metrics on the optimization process and outcomes.
	(4b) Pace layers	Changes in wellbeing are slow but AI optimization cycles are fast, making it difficult to optimize for wellbeing.

3.2.1. Challenges related to modeling wellbeing

Modeling wellbeing is crucial for designing AI systems that effectively promote human flourishing. However, it presents significant challenges due to the multifaceted nature of wellbeing and the complex sociotechnical contexts in which AI systems operate. This section explores two key challenges in modeling wellbeing: 1) identifying relevant wellbeing dimensions and 2) attributing causal relationships between AI system components and fluctuations in wellbeing.

Identifying relevant wellbeing dimensions

Since the rise of the Positive Psychology movement that gained popularity around the turn of the millennium (Seligman, 2019) scientific interest in wellbeing has proliferated. Wellbeing is a multifaceted, rich, and dynamic phenomenon, and as such, there are many definitions of it in both scholarly and public contexts (e.g., policymaking). Each definition pertains to different aspects of wellbeing. For instance, the WHO defines health as "a state of complete physical, mental, and social wellbeing, and not merely the absence of disease or infirmity." (Callahan, 1973) There are also scholarly traditions that call wellbeing "Happiness" which can be distinguished in episodic ("I feel happy") and attributed happiness ("I am a happy person")(Veenhoven, 2014). These definitions are akin to *hedonic* and *eudaimonic* understandings of wellbeing(Ryan & Deci, 2001)—which are the two categories typically used to describe wellbeing research.

Contemporary hedonic traditions tend to focus on the degree to which people experience positive and negative emotions and how satisfied they are with their life. Two reviews argue that the most prominent theoretical approach in this tradition is the tripartite model of Diener (Cooke et al., 2016; Linton et al., 2016). This model is typically measured in terms of "life satisfaction" (Diener et al., 1985) and the presence of positive and absence of negative emotions (Watson, Clark, & Tellegen, 1988).

In contrast, eudaimonic wellbeing is based on the pursuit of virtue, striving to become the best version of oneself and developing one's personal strengths (Deci & Ryan, 2008a). Psychologists investigating this phenomenon tend to define it in a multifaceted way, such as Ryff's Psychological Wellbeing (PWB) scale (Ryff, 1989; Ryff & Keyes, 1995) and Seligman's PERMA model (Seligman, 2010). These theories overlap to a great extent and encompass facets such as positive relations, meaning in life, and personal growth.

Further, there is the tradition of Quality of Life (QoL), which is often used interchangeably with wellbeing. However, the literature on the subject

often pertains more to social aspects of the phenomenon and, for example, situations towards end of life, living with a disability, or life after a clinical visit (Bakas et al., 2012; Felce & Perry, 1995; Moons, Budts, & De Geest, 2006).

A final related concept is wellness, which emphasizes a holistic lifestyle (e.g., nurturing emotional and spiritual intelligence) (Corbin & Pangrazi, 2001). In identifying the relevant dimensions of wellbeing, a designer could also investigate non-Western traditions of defining wellbeing such as Ubuntu (Hailey, 2008), Ikigai (Sone et al., 2008), and Gross National Happiness (GNH) (Ura, Alkire, & Zangmo, 2012).

Researchers criticizing the Western viewpoint in wellbeing science also argue that current mainstream approaches tend to be too contextualized (J. Mead, Fisher, & Kemp, 2021). Historically, psychological research on wellbeing has focused on psychometrics in order to statistically validate generalizable dimensions of wellbeing (Searle, Pykett, & Alfaro-Simmonds, 2021). However, this approach does not necessarily translate to actionable insights. Recognizing this, scholars have developed domain-specific theories of wellbeing for areas like work (Clifton & Harter, 2021) and education (Konu & Rimpelä, 2002).

Contrary to global conceptions of wellbeing, domain-specific theories focus more on aspects that are prevalent in that domain which may not translate to other domains. By contextualizing wellbeing research, the findings become more directly relevant for decision-making such as in policy-making. For example, they can reveal nuances around how wellbeing manifests with a given domain such as stress experience around work relations. However, increased specificity comes at the cost of reduced situational-consistency. This shift underscores the importance of balancing context-sensitive perspectives with more generalized insights that hold over time and place. Both remain indispensable for a comprehensive understanding of wellbeing.

The concept of digital wellbeing illustrates a similar balance. Digital wellbeing being refers to wellbeing effects that typically occur in a digital context—i.e., while gaming, surfing the web, or interacting on social media. It goes beyond the mere time spent online, focusing on how digital engagement affects daily life and emotional wellbeing. It is about achieving a harmonious interaction with technology, where the benefits are maximized, and drawbacks such as loss of control are minimized. Furthermore, the IEEE 7010-7020 (IEEE SA, 2020) initiatives represent a pioneering effort to establish a comprehensive understanding of wellbeing within the realm of AI. Other researchers have explored the relationship between AI and

community wellbeing (Musikanski et al., 2020) investigating the development of wellbeing metrics, community-centric AI, and applying AI to enhance community wellbeing.

Wellbeing, as highlighted in various perspectives, presents designers of AI systems with a range of dimensions and paradigms to consider. The relevance of wellbeing aspects varies by context; for example, physical health may be central in diet apps, while social media platforms might emphasize social connection. This complexity in people's interactions with technology is highly important for designers to consider. It requires careful, context-specific selection of the most relevant wellbeing paradigms and domains. Currently, there is no agreed-upon process for determining the which wellbeing dimensions to prioritize for a given system. This complexity in people's interactions with technology, being situational and dynamic, is reflective of findings in the literature (Vanden Abeele, 2021). As AI systems become increasingly entangled with daily routines, modeling their impacts on wellbeing becomes difficult. Thus, selecting appropriate theoretical dimensions is crucial yet challenging. Designers, therefore, face the complex task of modeling wellbeing in AI systems within evolving sociotechnical landscapes, a challenge we will explore further.

Modeling and attributing wellbeing changes

For AI to effectively promote wellbeing, it requires a deep understanding of what actions and strategies contribute to this goal. This, in turn, requires a contextual model of wellbeing that allows the designer to attribute fluctuations to specific features of the system. Contextual models of wellbeing explain how various aspects of wellbeing manifest situationally and connect to components of a given system. The discussion below focuses on the challenges in developing contextual wellbeing models and attributing wellbeing fluctuations to components of that model, starting with the former.

Considering the complexity of AI systems and contexts, it is crucial to develop an understanding of how different aspects of wellbeing relate to specific features or components of a particular context. Quantitative measures enables researchers to objectively examine explanations and predictions from conceptual models addressing the determinants and outcomes of wellbeing (Diener, 2019). However, developing such models poses significant challenges. The relationships between various causal factors and wellbeing remain unclear, complicated by bidirectional and nonlinear effects. For example, while good sleep benefits mental health, improved mental health also leads to better sleep (A. J. Scott, Webb, Martyn-St James, Rowse, & Weich, 2021). Similarly, the connection between mental health and social

media use is complicated: it is unclear whether people turn to social media when they are already struggling with mental health issues, or whether social media itself can contribute to these issues (Coyne, Rogers, Zurcher, Stockdale, & Booth, 2020; Hjetland, Schønning, Aasan, Hella, & Skogen, 2021).

The open questions around the relation between specific media and wellbeing have implications for the deployment of AI technologies aimed at promoting wellbeing (Johannes, Dienlin, Bakhshi, & Przybylski, 2022). Specifically, they highlight the need to consider the long-term impacts of these technologies on individuals' psychological health and wellbeing—e.g., watching Netflix may be conducive to wellbeing in the current moment but how may it shape effects over time? Thus, to design Positive AI, it is necessary to have a thorough understanding of the relationships between wellbeing and its various antecedents. Without this understanding, it is challenging to measure the impact of interventions and determine their effectiveness. The real world is characterized by complex, interconnected patterns that can make it difficult to attribute changes in wellbeing to a single event or intervention (Fokkinga, Desmet, & Hekkert, 2020). This is particularly true when dealing with "narrow" AI systems, such as recommendation algorithms that suggest products a user might like based on their past preferences. These systems may excel at completing specific, limited tasks but struggle to account for the full range of factors that can impact wellbeing in the real world. For example, a product recommendation system would not consider how using that product might affect a user's sleep, relationships, or long-term wellbeing.

This modeling poses difficulties not just due to AI system opacity (Gabriel, 2020), but because any platform comprises only a fraction of a person's broader life experience. For example, changes in the dietary practices of a teenager may be a result of the content of their Instagram feed (e.g., only images of people with a specific body type) or because of some other event that occurred in their life (e.g., a breakup or the start of a new fitness program). In order to establish causality, wellbeing has to be measured in a manner that considers the complexity of the context in which the system is deployed. This poses a challenge because existing wellbeing assessments and platforms are not designed to consider the complexity of people's lives and experiences that extend beyond the platform. For instance, how might Instagram account for life events that occur outside of its platform and influence a user's wellbeing? And how would it determine which of those external factors are most relevant and impactful? The question of whether Instagram should be held accountable for the wider impact of its user

interactions on wellbeing is a matter of ongoing debate. However, for the platform to effectively influence wellbeing, it is imperative that it addresses these interaction effects in some capacity.

In summary, developing contextual models that effectively attribute fluctuations in wellbeing to AI systems poses profound challenges. The interconnected relationships between wellbeing and other life factors resist straightforward causal analysis. This difficulty intensifies for narrow AI platforms, as they comprise limited slices of broader human experience. Although difficult, advancing more contextualized and systemic modeling methodologies promises significant progress toward the goals of Positive AI.

3.2.2. Challenges related to assessing wellbeing

Wellbeing is suited for AI optimization given its history of measurement (Stray, 2020). However, to measure wellbeing in context, we need to model wellbeing in context. For this, we need qualitative inputs to understand subjective personal and community experiences. Translating qualitative insights into quantitative metrics usable for optimization is non-trivial. Here, we discuss challenges related to effective assessment, which requires contextualization and bridging gaps between individual perspectives and system-level scales.

Contextually operationalizing wellbeing

As previously discussed, in order to attribute fluctuations in wellbeing to components of a given system, wellbeing must be effectively modeled and measured. While we have explored the theoretical challenges associated with this task, there are also methodological aspects that need to be addressed. This includes the operationalization of chosen wellbeing dimensions.

Wellbeing has traditionally been measured in field of psychology, but since the last decade, there has been an increased interest in also measuring wellbeing for other purposes, such as policymaking (Frijters & Krekel, 2021). The assessment of wellbeing is often done using qualitative surveys and interviews (Alexandrova, 2012). It is important that the assessment instruments employed in these studies are validated, possess temporal stability, and demonstrate cross-situational consistency (Diener & Michalos, 2009). However, literature suggests that these "off-the shelf" wellbeing assessment instruments may not be readily applicable in the context of novel technologies that rapidly change factors influencing wellbeing, such as social media (Kross et al., 2021) and AI (Stray et al., 2023). Traditional instruments aiming for consistency over time and place may be incompatible

with contexts that transform quickly and substantially. This also means that present modalities of assessment (e.g., surveys and interviews) may not be sufficient for these emerging contexts (Stray et al., 2023).

Given these limitations, there is a need for measures that are better suited to assess wellbeing in the context of rapidly changing technologies and situations (Vanden Abeele, 2021). Context-sensitive measures, designed to adapt to the specific needs and context of the individuals being assessed (Loveridge, Sallu, Pesha, & R Marshall, 2020; van der Maden, Lomas, & Hekkert, 2023), provide a promising alternative and have been suggested to be more suitable for the operationalization of values (including wellbeing) in concrete applications (Liscio et al., 2021). The value inference process outlined by Liscio et al. (2023) highlights the importance of identifying context-specific values as a crucial step in aligning AI agents with human values. Their methodology demonstrates an approach to elicit relevant, contextualized values that could inform the development of context-sensitive wellbeing measures. These measures allow for a more tailored assessment of wellbeing, as they can be regularly updated to reflect the changing needs of the community. It should be noted that this adaptability poses tradeoffs regarding validation and stability over time and contexts. How to reconcile the discrepancy between the need for sensitive, customized measures and generalizable instruments remains an open question warranting further investigation.

Finally, to develop context-sensitive measures of wellbeing, it is crucial to engage regularly with the community to scrutinize and update the instruments used to assess contextual wellbeing. The literature has identified the importance of community engagement to implement ethical frameworks (Morley, Elhalal, et al., 2021) and articulate shared values in AI systems (Sanderson et al., 2023), though sources acknowledge there are gaps in best practices for stakeholder participation (Sadek et al., 2023a). While human-centered design methods may provide useful directions for establishing community engagement, they need to be adapted to the particularities of AI systems. Human-centered design excels at developing a deep understanding of contexts, maintaining community participation, and understanding individual and community needs. In conclusion, contextualizing wellbeing measures involves navigating consistency-sensitive tensions and engagement complexities. However, even well-constructed contextual measures struggle to inform AI optimization unless translated to system-level data. This underscores the pivotal challenge of connecting qualitative insights with large-scale, quantative metrics for AI alignment.

Translating qualitative experiences into system metrics

Wellbeing is a highly personal experience that typically requires individuals to report on their own experiences in order to be measured (Diener & Michalos, 2009; Linton et al., 2016). As such, qualitative methods such as 1-to-1 interviews, focus groups, and ecological momentary assessments, are often expected to provide more contextually actionable data. However, currently, there is no agreed-upon process for translating small-scale activities of this nature into large-scale optimization metrics (McGregor, Camfield, & Coulthard, 2015). Instead, easy-to-collect but incomplete or inaccurate metrics are used—such as hours spent on a social media platform to measure satisfaction with content—which do not fully capture users' experiences and overlook harmful consequences to their wellbeing (Thomas & Uminsky, 2020). Conversely, connecting qualitative research on wellbeing with large-scale optimization methods could not only help ensure that measures of wellbeing are well-aligned with human experiences, but also help identify areas in which measures of wellbeing can be improved (Camfield, 2016). Aside from issues of scalability, there exists a discrepancy between the types of metrics suitable for on-platform measurement versus those typical for wellbeing assessment.

Whereas self-reporting suggests to be the best way of measuring wellbeing, behavioral data collection is the default method for on-platform optimization. However, behavioral metrics cannot reliably measure wellbeing since research on the relationship between behavioral and self-reported wellbeing measures is limited and inconclusive (Dang, King, & Inzlicht, 2020). Currently, it is unclear whether behavioral metrics can replace self-report measurements. Relying solely on self-reporting may, however, negatively impact the user experience, as users should not be bombarded with wellbeing questions upon engaging with a platform. This presents a challenge for designers and design researchers: how can platforms facilitate user feedback to collect accurate and scalable wellbeing data? This non-trivial challenge has also been acknowledged by other researchers (e.g., Steur & Seiter, 2021).

In conclusion, effectively integrating wellbeing requires translating between qualitative experiences and system metrics. While qualitative insights like self-reports capture personal experiences vital for alignment, convenient behavioral metrics dominate on-platform data collection. Absent mechanisms to translate small-scale activities into optimization inputs, AI risks misrepresenting user needs. Progress necessitates the development of methods that bridge this divide —continuously engaging individuals and communities while surfacing priorities at a systemic level. However, creating participatory channels poses immense practical difficulties

around incentivization, standardization, and scalable synthesis. Though an open challenge, pioneering such participatory architectures in ways that meaningfully empower stakeholders promises to actualize AI's potential for responsibly nurturing human flourishing.

3.2.3. Challenge related to designing for wellbeing

We have discussed the difficulties in conceptualizing and operationalizing wellbeing for AI systems. We have also touched on the need to optimize across individual, community and societal levels of wellbeing. A further question is: given a concept and operationalization of wellbeing, what actions can an AI system take to positively impact wellbeing? And how may we design such actions?

Methodological challenge of designing (AI) actions to promote wellbeing

Designing AI is an incredibly challenging task for designers (Sadek et al., 2023b; Yang et al., 2020). Before tackling the design of AI that promotes wellbeing, it is important to understand the difficulties that designers face when working with any AI-based system in general. Currently, the communication gap between designers, developers, and end-users (Yang et al., 2020; Yu et al., 2020) causes an "AI support vacuum" (Abaza, 2021) where AI neither supports stakeholders nor is supported by them. Despite calls for the more interdisciplinary design of AI-based systems (Harbers & Overdiek, 2022; West, Whittaker, & Crawford, 2019), there is a lack of 'translational work' in current interventions that aim to support this collaborative design (R. Wong et al., 2022). Aside from communication-based challenges, without proper training, it is difficult for designers to understand how to ideate, design and prototype for AI-based systems (van Allen, 2018). It is important to mention and consider these challenges before examining the extra difficulties that a focus on wellbeing might present.

Following the earlier discussion of adopting a systemic perspective, it is currently unclear whether designing Positive AI requires changes to the interface, algorithms, content moderation policies, business models, or otherwise defined components. Interventions could even extend beyond the system itself. This uncertainty stems from a lack of examples and established methods for putting wellbeing at the core of AI design. For example, current strategies for promoting wellbeing through ChatGPT remain largely undefined and untested. Designing Positive AI requires fundamentally

rethinking how we approach design to focus on the continuous measurement of wellbeing and alignment with human values. There are well-established frameworks for designing with wellbeing in mind, such as Positive Design (Desmet & Pohlmeyer, 2013) and Positive Computing (Calvo & Peters, 2014). These approaches acknowledge the need to consider the impacts on wellbeing, but they were not developed with the unique complexities of AI systems (such as system opacity, unpredictability, and scalability) in mind.

In addition to design methods, alignment approaches like Constitutional AI (Bai et al., 2022), Inverse Reinforcement Learning (IRL) (Arora & Doshi, 2021; A. Y. Ng & Russell, 2000) and Contestable AI (Alfrink, Keller, Kortuem, & Doorn, 2022) also aim to ensure AI systems remain human-centered and aligned with human values. While these methods primarily focus on modifying the AI system's internal processes and decision-making mechanisms, they highlight the importance of incorporating human values and wellbeing considerations directly into AI systems. By combining the strengths of these alignment approaches with design methods, we can develop a more holistic and effective AI alignment strategy that addresses user experience, interaction, and overall design, which are crucial for affecting wellbeing. For instance, some platforms, like YouTube and Twitter, have made small changes intended to benefit wellbeing, such as removing dislike counts and allowing users to 'unmention' themselves from conversations. Whether these are cases of 'ethics washing' (Floridi & Cowls, 2019) or have actual benefit to wellbeing is currently unknown to the public and academia, as research on their effects is not shared publicly.

In this vein, Stray (2020) discuss Facebook's MSI metrics and YouTube's satisfaction metrics as cases of AI optimization aligned to wellbeing. However, they criticize that both companies did not involve or get feedback from the people affected by their AI changes. The absence of public assessment and unclear impact on broader aspects of wellbeing, such as social connectedness or life satisfaction, particularly in diverse communities, underscores a crucial shortcoming. This lack of detailed information and engagement hinders a comprehensive understanding of the effectiveness of these interventions in truly aligning AI with community wellbeing. This is a broader concern in the value-sensitive design (VSD) of AI. That is, as discussed in the background section, while VSD methods support the identification and embodiment of values for AI well, they lack support in assessing 'realized' values—i.e., whether the designed outcomes in fact achieve the intended effect on said value (Sadek et al., 2023b). This is particularly important for Positive AI, because the core premise lies in empirically confirming that systems positively impact human wellbeing.

Without closing the loop between intended and actualized outcomes, the benefits of proposed interventions remain theoretical.

To recapitulate, designing AI actions that support wellbeing is a complex challenge that requires a combination of design methods and alignment approaches. Successful regulation and alignment requires internal diversity matching external complexity. Without design methods and alignment approaches considering the full range of variables shaping wellbeing, AI systems will struggle to effectively promote flourishing While existing design methods provide guidance, further research and development is needed. By focusing on continuous measurement of wellbeing, alignment with human values, and incorporating aspects of both design methods and alignment approaches, we can work towards creating Positive AI systems that not only avoid harm but actively promote human flourishing.

3.2.4. Challenges related to optimizing wellbeing

Optimization tradeoffs

Maximizing user engagement to drive revenue is a central optimization challenge for most platforms, where varied metrics like views, likes, and shares track user interaction with content. Using multiple metrics allows optimization algorithms to take a multi-objective approach to personalize content for optimal user engagement (Trunfio & Rossi, 2021). However, balancing these objectives already involves tricky tradeoffs (Thorburn, 2022). For example, there is tension between showing novel content to pique interest and only showing content matched to the user's interests to avoid disengagement (Lu, Dumitrache, & Graus, 2020). While some novelty draws users in, too much risks boring them with irrelevant content. Now, considering wellbeing optimization makes this process even more complex.

Firstly, it is difficult to determine which facet of wellbeing should be prioritized—both in the moment and over time. The various facets of wellbeing often compete with one another. For example, social media platforms must balance users' needs for social connection and personal autonomy. Promoting social interaction may support wellbeing by facilitating relationships, but it could also infringe on users' freedom to choose their own activities. Additionally, an individual's priorities may shift over time. What enhances wellbeing in the short-term may differ from long-term needs (e.g. enjoying frequent social activities when moving to a new city versus after settling in). This complexity requires nuanced techniques that can account for tradeoffs between competing needs and evolving individual priorities.

Secondly, because of these optimization tradeoffs there is no optimal

solution. This is a common issue in environmental sciences who face complex tradeoffs between, for example, biodiversity and human wellbeing (Daw et al., 2015; McShane et al., 2011). This also goes for optimizing for just human wellbeing in the sense that there is no "right" outcome in balancing, for example, the wellbeing of an individual over that of a family (who are likely to value different things). Here the notions of requisite variety and satisficing become important again. That is, as platforms are faced with these complex tradeoffs they should develop many different strategies to deal with problems as they come up and take small incremental steps to deal with tradeoffs over time.

Thirdly, prioritizing one wellbeing facet over others may yield unintended consequences—beyond just direct tradeoffs. Focusing narrowly on a single aspect of wellbeing can backfire and undermine that very facet over time, given the complexity and interconnectivity of wellbeing. For instance, features meant to enhance social connectedness could hamper autonomy or other unforeseen needs. Wellbeing's multidimensional nature means myopic solutions risk negative ripple effects from complex causal interrelations that remain poorly understood. Therefore, maintaining broad sensitivity to these complex interactions is critical. By regularly reassessing for subtle harms and tightening feedback loops, platforms can progressively identify and resolve unintentional side effects. An open, responsive systems perspective allows more complete understanding to emerge gradually from ongoing learning. Overall vigilance to complexity and interconnectedness may better serve wellbeing than rigid prioritization of singular facets.

Therefore, when designing AI systems for wellbeing, the models guiding decisions should be continually reassessed through collaboration with relevant stakeholders. This allows for adaptive alignment as understandings of wellbeing, design contexts, and community needs evolve over time. Continual engagement enables updating system priorities to restore balance across wellbeing dimensions. However, even with continual reassessment and stakeholder collaboration, another fundamental challenge persists—the differential pace of change between wellbeing and AI optimization.

Fundamental challenge of pace

Changes in wellbeing typically manifest over longer periods of time. For instance, life satisfaction is not expected to fluctuate dramatically from week to week (Pavot, 2014). While it can be argued that this is due to our assessment instruments (Boschman, Nieuwenhuijsen, & Sluiter, 2018), the argument remains that this is in stark contrast to the pace of AI optimization and, for instance, media consumption, which are much faster. This can make

it difficult to reconcile the two, as the wellbeing effects of AI actions should inform the system's optimization cycles. Note here that system optimization cycles here refer to both analog (e.g., managerial, and designerly) and digital (e.g., algorithmic news feed recommendations) optimization cycles.

In other words, from the perspective of the optimization algorithm, the best it can do is optimize for hedonic wellbeing (momentary pleasure), while the goal is to design AI that supports eudaimonic wellbeing too (long-term wellbeing). An AI system can optimize for what feels good in the moment but not for what is good for you in the future. For example, watching "just one more episode" may be desirable in the moment (hedonic) but regretted the next morning when you feel tired during an important meeting (eudaimonic). Alternatively, from a managerial optimization perspective, linking executive goals to user wellbeing metrics may show progress only after longer periods than quarterly reporting cycles. Yet leaders frequently anticipate prompt, measurable outcomes that can be effectively communicated during shareholder meetings, reflecting a preference for quick, tangible results. This pace mismatch risks reactive changes before initiatives fully play out—such as prematurely disbanding a wellbeing taskforce. Evidently, there is a need to reconcile these pace layers somehow.

Here, we can look to Stewart Brand's theory of pace layering (Brand, 2018), which suggests that different systems evolve at different speeds (see Figure 3.2). The pace of platform change far outpaces that of wellbeing. This is a fundamental constraint: the pace of AI optimization processes and the timeline for observable changes in human wellbeing are mismatched, and the rate of these dynamics in either domain is unlikely to significantly change in the future. Acknowledging and making explicit that there are in fact different components in a system that evolve at different paces is essential as we should look for ways to bridge or translate across layers. That is, we need an intermediary layer for measurement—for which qualitative methods may be well-suited. In the example given before, Netflix may choose to incorporate qualitative feedback from users regarding their overall satisfaction with their viewing habits, including how it affects their daily lives, sleep patterns, and long-term goals—bridging the gap between behavior and AI actions.

3.3. Discussion

This paper aimed to outline key challenges designers face when developing AI systems to actively promote human wellbeing, termed Positive AI. It makes two main contributions. First, it proposed adopting a cybernetic, systemic perspective for conceptualizing and addressing these alignment

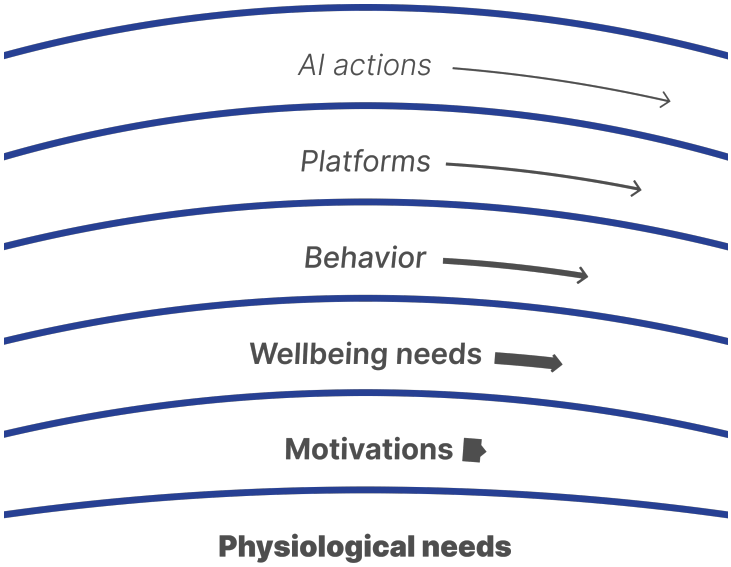


Figure 3.2: Different systems operate at different paces, adapted from Stephen (2021)

challenges. This viewpoint emphasizes the sociotechnical nature of AI systems, considering potential interventions across multiple system levels. Second, it organized the complex issue of designing Positive AI into four main categories of challenges: 1) modeling wellbeing, 2) assessing wellbeing, 3) designing for wellbeing, and 4) optimizing for wellbeing. These categories provide structure for mapping relevant problems within a systemic framework geared towards continuous improvement through stakeholder participation.

To recapitulate the challenges, designing Positive AI is facing substantial gaps regarding our knowledge on how to do it. The abundance of theoretical wellbeing paradigms makes it difficult for designers to get started on modelling wellbeing within a given context. This leads to issues in measuring wellbeing and attributing changes across different scales, from individuals to communities to sociotechnical systems. Further, the complexity of wellbeing introduces optimization tradeoffs, raising questions around balancing competing wellbeing needs between individuals and groups over time. While these present methodological and technical difficulties, they are fundamentally design challenges requiring creative solutions.

Because of these challenges, in some cases, platforms may opt for simpler derivatives of wellbeing (Pan, Bhatia, & Steinhardt, 2022), or designs that merely mitigate illbeing or improve aesthetics, rather than investing resources

in optimizing for wellbeing itself. However, the impending mental health crisis (Organization, n.d.) and the transformative impact GenAI (Chui et al., 2023) make it all the more urgent to move towards Positive AI that systematically prioritizes wellbeing. This requires more than just a focus on mitigating harm; it is about capitalizing on each platform's potential to foster wellbeing.

The key takeaway is that adopting a cybernetic perspective that places wellbeing assessment at the core can guide this process. It emphasizes the need for design researchers to continuously re-examine contextual models of wellbeing. This requires engaging relevant stakeholders, such as end-users and developers. By framing wellbeing as an ongoing conversation, we can iteratively refine models and measurements as contexts evolve through stakeholder participation. This reflexive, adaptive approach allows designers to navigate complexity and uncertainty when developing AI systems aimed at fostering human flourishing.

Wellbeing as an objective for AI optimization

The concept of optimizing metrics wellbeing in AI systems emerges as a response to the limitations of traditional metrics like profit or efficiency (Schiff, Ayesh, et al., 2020; Stray, 2020). While financial returns have been the primary focus for many companies, this singular pursuit may lead to broader societal harms (Virokannas, Liuski, & Kuronen, 2020). Might wellbeing as a guiding principle for AI design serve as a compelling alternative? Wellbeing, with its quantifiable and multidimensional nature, encompasses various aspects of human life which may make it a more suitable optimization goal. It allows for concrete optimization metrics that are more aligned with human-centered values, potentially reversing the trends of societal harm caused by narrowly focused objectives.

Yet, it seems impossible (and ill-advised) to reduce wellbeing to a single metric. Instead, wellbeing should serve as a multidimensional guide for value-based decisions and a comprehensive principle for moral choices.

Despite arguments for optimizing AI systems for wellbeing, companies may still opt for simpler derivatives due to factors like potential public relations issues, unclear financial rewards, and risks of losing competitive advantages. For example, companies may hesitate to invest in research on the effects of their products on wellbeing, as negative findings could result in bad publicity as exemplified by the 2018 'Techlash' at Facebook for instance (Hemphill, 2019). Further, the benefits of prioritizing user wellbeing over profits are currently ambiguous for technology companies. Specifically, companies grapple with uncertainty about whether prioritizing

user wellbeing over profits, which often involves focusing more on 'doing good' rather than just 'preventing harm,' will yield tangible benefits (Morley, Kinsey, et al., 2021). This ambiguity is rooted in the difficulty of quantifying the return on investment for ethical AI practices that emphasize proactive welfare measures over mere harm avoidance. However, gaining clarity on this trade-off is impeded by corporate reluctance toward transparency and third-party algorithm audits, which are seen as jeopardizing competitive advantage (Stray & Hadfield, n.d.; Stray et al., 2023). This uncertainty, coupled with the perceived risks of optimizing for wellbeing, disincentivizes companies from allocating resources to human-centered design interventions. Overcoming these barriers will require establishing transparent accountability mechanisms, alternative business models not reliant on exploitation, and fostering an ethical culture recognizing that benefiting humanity and profits can be compatible (Di Vaio, Palladino, Hassan, & Escobar, 2020).

This reluctance highlights intricate tensions between public perception, economic incentives, and ethical duties facing corporate decision-makers. Such challenges of power, as also publicly displayed during the OpenAI-Altman debacle late 2023 (Ulanoff, 2023), are highly relevant for Positive AI. These second-order challenges determine system-level goals, shaping whether wellbeing optimization occurs. If companies lack motivation to prioritize wellbeing, alignment is unlikely. Wellbeing as an overarching principle for AI alignment is promising but faces real-world obstacles regarding corporate priorities. While prominent voices endorse human flourishing as the goal, transparency and accountability mechanisms appear necessary to actualize this vision (Morley, Elhalal, et al., 2021).

Future opportunities

AI benchmarking is a popular method for evaluating the capabilities of large language models (LLMs). As AI benchmarking matures and as AI permeates more aspects of life, more sophisticated will be benchmarks required (Burnell et al., 2023). For example, benchmarks for LLM qualities like toxicity are now widely used (Lynch, 2023). Carefully crafted wellbeing metrics could serve as a mechanism for academics and others to indirectly optimize AI systems such as ChatGPT. This is because benchmarks are used to compare different models; as a result, low performance on a benchmark can motivate improvement.

Human-centered designers may help attune evaluation methods such as benchmarking better to actual human experiences and ensure that optimization metrics align with these experiences.

3.3.1. Limitations and final remarks

While this paper aimed to provide an overview of key challenges in designing Positive AI, there are inherent limitations in addressing such a complex, transdisciplinary topic. For instance, environmental sustainability is not a core focus of the article, yet it warrants mention given the intricate relationship between sustainability and human wellbeing. As environmental crises increasingly threaten flourishing across communities, sustainability is being recognized as fundamental to comprehensive models of wellbeing (Kjell, 2011; O'Mahony, 2022). That is, wellbeing means more than human wellbeing. Meanwhile, the energy-intensive nature of AI systems presents sustainability challenges (Vinuesa et al., 2020). This surfaces an alignment tension between AI benefits and potential unintended harms. Additionally, many relevant issues, from philosophy of technology to data ethics, could only be briefly touched upon given the practical design focus of this paper, but present key areas for further investigation.

Nonetheless, this work aims to spark discussion and research at the intersection of AI, wellbeing, and design. Further interdisciplinary collaboration building on these ideas will develop more pluralistic perspectives on Positive AI. The goal of this work was to identify key challenges that must be addressed by designers in order to develop AI systems aligned with human wellbeing. As the fourth industrial revolution is well on its way, "the time of reckoning for artificial intelligence is now." (Ozmen Garibay et al., 2023)

4

Case study: My Wellness Check

Bringing Positive AI from theory to reality requires overcoming practical barriers through iterative, human-centered approaches. This chapter illustrates such challenges through the large-scale empirical case study of *My Wellness Check*, a cybernetic system designed to support student and staff wellbeing during COVID-19. Spanning over 20,000 participants across two years, the project included seven assessment iterations. It reveals barriers empirically manifesting across key processes, while showcasing grounded responses like continual community engagement. By thoroughly examining the iterative cycles of development and evaluation undertaken to refine this sociotechnical wellbeing tool, the chapter distills essential insights. These practical learnings directly inspired the development of a human-centered design method for Positive AI proposed in the next chapter. This extensive case study showcases an adaptive methodology beginning to actualize Positive AI's potential to foster flourishing.

This chapter was previously published as "van der Maden, W., Lomas, D., & Hekkert, P. (2023). A framework for designing AI systems that support community wellbeing. *Frontiers in Psychology*, 13, 1011883. https://doi.org/10.3389/fpsyg.2022.1011883"

Designing artificial intelligence (AI) to support health and wellbeing is an important and broad challenge for technologists, designers, and policymakers. Drawing upon theories of AI and cybernetics, we offer a design framework for designing intelligent systems to optimize human wellbeing. Our framework focuses on the production of wellbeing information feedback loops in complex community settings.

The basis for our discussion is the community-led design of My Wellness Check, an intelligent system that supported the mental health and wellbeing needs of university students and staff during the COVID-19 pandemic. Our system was designed to create an intelligent feedback loop to assess community wellbeing needs and to inform community action. This article provides an overview of our longitudinal assessment of students and staff wellbeing (n = 20,311) across two years of the COVID-19 pandemic. We further share the results of a controlled experiment (n = 1,719) demonstrating the enhanced sensitivity and user experience of our context-sensitive wellbeing assessment.

Our approach to designing "AI for community wellbeing," may generalize to the systematic improvement of human wellbeing in other human-computer systems for large-scale governance (e.g., schools, businesses, NGOs, platforms). The two main contributions are: 1) showcasing a simple way to draw from AI theory to produce more intelligent human systems, and 2) introducing a human-centered, community-led approach that may be beneficial to the field of AI.

4.1. Introduction

Artificial Intelligence (AI) is transforming our global society, from creative industries to healthcare. However, as the negative impacts of AI become more apparent—whether regarding employment automation (Cugurullo & Acheampong, 2023) or social media's harm on wellbeing (Kross et al., 2021)—action is needed.

As AI is increasingly viewed as a potential solution to many large challenges, a variety of organizations have investigated how AI might support the mental health and wellbeing needs of their stakeholders (employees, customers, students, etc.). However, rather than seeing "AI for wellbeing" as a specialized interest to the mental health community, we argue that all ethical AI systems have an implicit objective to enhance human wellbeing. For instance, according to a IEEE standards review, "by aligning the creation of [AI] with the values of its users and society we can prioritize the increase of human wellbeing as our metric for progress in the algorithmic age." (Shahriari & Shahriari, 2017) While optimizing wellbeing may be a key goal

of AI systems aligned to human values, there remain many challenges to assessing or measuring human wellbeing at scale.

What do we mean when we refer to AI systems? The European Commission defines AI as "systems that display intelligent behavior by analyzing their environment and taking actions—with some degree of autonomy—to achieve specific goals." (Commission, 2019) For a second definition of AI, we look to a popular AI textbook that defines the field of artificial intelligence as the "study and design of intelligent agents." They define an agent as "anything that can be viewed as perceiving its environment through sensors and acting upon that environment through actuators." Finally, the intelligence of an agent is the "ability to select an action that is expected to maximize [a] performance measure...an agent that is assigned an explicit "goal function" is considered more intelligent if it consistently takes actions that successfully maximize its programmed goal function." (Russell & Norvig, 2022, p.58) Based on these definitions, AI systems require the ability to sense their environment, the ability to act on their environment, the ability to measure an explicit goal state in the environment (i.e., a performance measure or objective function), and the ability to use sense data to chose actions likely to improve that performance measure. These criteria lead to a proposed framework for designing AI systems (Table 4.1).

Instead of using AI tools (like chatbots) to support mental health and wellbeing, this article focuses on using theories developed by AI researchers to better understand how large-scale systems can be designed to better support wellbeing outcomes. Specifically, the theory that all intelligent systems are defined by their ability to assess system outcomes (performance measures) and choose subsequent actions in order to optimize those measures. This theory suggests that it may be inevitable that future wellbeing-aligned AI systems will necessarily need mechanisms for assessing human wellbeing. This paper provides a demonstration of developing context-sensitive wellbeing assessments that may inform the design of future AI system assessments of wellbeing.

This paper presents the community-led design of *My Wellness Check*: an intelligent system that measures human wellbeing in order to optimize and support the needs of university students and staff during the COVID-19 pandemic. Designed in collaboration with students, staff, and mental health professionals, the My Wellness Check system provided a governing feedback loop capable of assessing community wellbeing needs and informing community action. Based on theoretically-derived factors of wellbeing as well as factors defined by community participants, My Wellness Check produced real-time insights into community wellbeing that were used to inform actions

Table 4.1: A Framework for Designing AI Systems, based on a definition of AI focused on the "ability to select an action that is expected to maximize [a] performance measure" (Russell & Norvig, 2022)

	Step	Description
1	Task Environment	Define the task environment in which the AI system will be used and the requirements for success in that environment
2	Performance Measures	Develop performance measures or objective functions that quantify the goals of the system.
3	Action space	Identify the set of possible actions that the AI system can take in the environment and the set of possible states that can result from those actions
4	Sensors	Define a set of features that can be used to describe the state of the sensed environment
5	Algorithms	Define a set of algorithms that can be used to map the features of the environment to the actions that the AI system can take, for the purpose of optimizing the objective function—where the algorithm need not be software (Gillespie, 2014)
6	Implementation	Implement the system within the constraints of the environment, the users, and other stakeholders—the designers should remain an integral part of the implementation procedure and monitor performance (Norman and Stappers, 2015)
7	Refine	Based on feedback from users and other stakeholders, refine the system as necessary to improve performance

at various levels of the university, from top administrators to individual students. We share data from a longitudinal deployment of My Wellness Check to nearly 30,000 students and staff across two years of the COVID-19 pandemic. To evaluate our system, we share the results of a controlled experiment comparing our community-led wellbeing assessment to other wellbeing assessments. This shows that our community-led designs generated greater predictive value and a significantly better user experience. While our results cannot serve as proof of efficacy for the performance of our entire

system, it does show the benefits of our community-led design process.

Schools, businesses, NGOs, social platforms and other large-scale governing systems may wish to systematically improve the wellbeing of the people they serve. Our work aims to provide insights that can generalize to these different contexts. Rather than designing a fully autonomous system (for example, a chatbot to help provide students with mental health recommendations), we focused on introducing an intelligent feedback loop to an existing sociotechnical system.

Paper overview

In the first part of this article, we provide an overview of the concept of "designing AI for Wellbeing" and review some related efforts. We then discuss several methods and ideas popular within the field of human-centered design, such as participatory design, community-led design, systems thinking, and cybernetic thinking to address some of the challenges of designing AI for Wellbeing.

In the second part of the paper, we describe the specific context and the design of My Wellness Check. We then present data from multiple assessments of wellbeing over the period of the pandemic. Following a description of the design of the system, we present the design and implementation of a controlled experiment to evaluate our context-sensitive assessment. Following the presentation of the results of this experiment, we then reflect upon our design framework and suggest opportunities for future research in the design of AI systems for Community Wellbeing.

4.2. Related work: designing AI for wellbeing

There is a small, but growing, body of work on the use of AI for wellbeing or mental health, much of which focuses on the use of AI for health monitoring and personalized health advice. Often these services are delivered through the use of virtual agents, chatbot, wearables and other Internet of Things (IoT) technologies (Shah et al., 2003, see review by). According to D'Alfonso (2020) the three main applications of AI in mental health are: 1) personal sensing or digital phenotyping; 2) natural language processing of clinical texts and social media content; and 3) chatbots, while another review found opportunities for AI in mental health mainly related to self-tracking and AI assisted data analysis (S. Graham et al., 2019).

A 2020 Designing Interactive Systems (DIS) workshop on wellbeing offered the following summary of the field: "Most human-computer interaction

(HCI) work on the exploration and support of mental wellbeing involves mobiles, sensors, and various on-line systems which focus on tracking users." (Sas, Whittaker, Dow, Forlizzi, & Zimmerman, 2014) This reflects a focus on user-centered solutions for wellbeing, where wellbeing is conceptualized as the concern of an individual person.

In this paper, we present an alternative design objective: to support the wellbeing of a community of people. The wellbeing of a community can be understood as a multidimensional set of values, including economic, social, and environmental values, that impact people in a community (Musikanski et al., 2020). One advantage of this approach is that it does not require tracking individuals over time, which poses more risks from a data privacy and security perspective. Individual tracking, when it reveals deficits in wellbeing, may be damaging to individual self-image and produce feelings of guilt or disappointment (Chan et al., 2018).

A 2019 review of HCI technologies for wellbeing proposes the following: "We argue for an ethical responsibility for researchers to design more innovative mental health technologies that leverage less the tracked data and more its understanding, reflection, and actionability for positive behavior change." (Sanches et al., 2019) By focusing on wellbeing at a community level (namely, the students and staff at a campus university), we can avoid data tracking issues and include diverse stakeholders that can assist with understanding the wellbeing data, reflecting upon it and formulating approaches for positive action.

As part of our community-focus, our work centers around two components of the system: assessments of community wellbeing needs and the design of interventions that target those needs. We are inspired by cybernetic theory to design our system to produce a wellbeing feedback loop that supports both top-down and bottom-up processes. This approach lends itself to our participatory and community-led design methods It also stands in contrast to the objective of developing an autonomous system that uses a black-box, algorithmic approach to intervene in the community. Finally, we use an iterative, longitudinal design approach that emphasizes improvements in the assessment of wellbeing and the processes taken to transform those assessments into action.

4.2.1. Cybernetics: a conceptual framework

Cybernetics has seen a resurgence of interest due to the increasing popularity of artificial intelligence and machine learning (Pangaro, 2017, 2021), and can be seen as its conceptual predecessor (Figure 4.1). Partially explaining this interest is the common (mis)conception that the purpose of artificial intelli-

gence is to replace human intelligence with computational intelligence—also called "AI thinking" (van der Maden et al., 2022). As artificial intelligence does typically focus primarily on computational systems, cybernetics offers a conceptual framework for understanding the design of systems that are capable of purposeful (intelligent) behavior—regardless of whether the systems involved are natural, artificial, or a mix of the two.

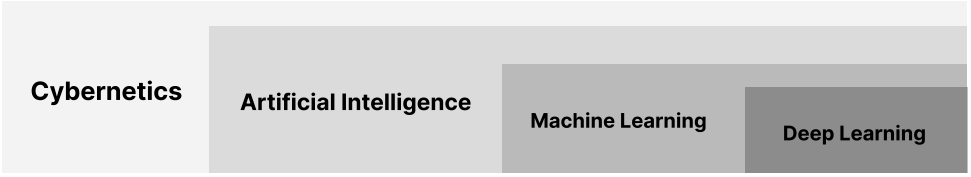


Figure 4.1: Representation of cybernetics as the conceptual predecessor of AI used with permission from J. D. Lomas, Patel, and Forlizzi (2021).

Cybernetics can be described as the interdisciplinary study of the design of governing systems, both human and machine, that use sensors and actuators to achieve a goal. The word cybernetics comes from the Greek word "kyvernitis" or "kubernetes," means "steersman" or "governor" (note that the verb "to govern" also comes from this Greek root). Cybernetics has been used to help design everything from robots to organizations. It has also been used to study human cognition and social interactions.

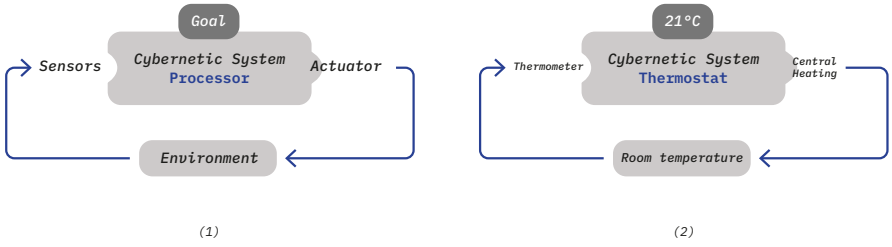


Figure 4.2: The schematic on the right (1) is an abstraction of a typical cybernetic system (Dubberly & Pangaro, 2007, adapted from). The schematic on the left (2) shows a typical example of a cybernetic system, a thermostat.

A cybernetic system is a system where feedback loops are used to control the behavior of the system. At its simplest, a cybernetic system consists of just three parts: a controller, a sensor, and an effector or actuator.

In a simple home thermostat, for example, the sensor is the temperature sensor, the effector is a switch that turns on a heater, and the controller is a mechanism that compares the sensor to a point set by the user. If the sensor value is below the set point, the controller turns on the heater, see Figure 4.2. If the sensor value is above the set point, the controller turns off the heater. More complex "smart thermostats" may have additional sensors (e.g., humidity, occupancy, etc.), and effectors (e.g., air conditioner, fan, etc.), and the controller may use a more sophisticated algorithm to determine when to turn devices on and off.

Cybernetic systems are not restricted to simple devices like thermostats, however. Cybernetic systems can be found in living organisms (e.g., the feedback loops that control blood sugar levels), in social systems (e.g., the feedback loops that govern the interactions between people), and in artificial intelligence systems (e.g., the feedback loops that allow a robot to learn from its mistakes). Cybernetics is closely related to the field of artificial intelligence. Both fields are concerned with the design of adaptive systems. A typical example is reinforcement learning, which is a machine learning method that uses rewards or punishment to train an agent to perceive and interpret its environment and take actions, see Figure 4.3. The fields also differ in their typical focus. For instance, cybernetics tends to be more concerned with understanding or designing feedback mechanisms that allow a system to govern its behavior, whereas artificial intelligence is more concerned with the design of algorithms that allow a system to learn from or adapt to its environment. Second, cybernetics tends to be more concerned with the design of natural systems (i.e., designing human

Figure 4.3: A schematic of a typical reinforcement learning algorithm from adapted Sutton and Barto (2018).

governing systems), while artificial intelligence is more concerned with the design of computational systems.

Practically speaking, cybernetics offers a viewpoint for designing intelligent systems for governance that include both computers and people (Dubberly & Pangaro, 2019; Glanville, 2009; Krippendorff, 2007, 2021; Sweeting, 2016, e.g.,). Replacing human intelligence with computational automation is often not desirable, largely due to the special capacities of human interactions. Instead, there is a need to design systems, both natural and computational, that work together to create more intelligent behavior (i.e., more able to achieve goals in an uncertain environment). Cybernetics provides a means for conceptually uniting humans and artificial systems While keeping "humans in the loop" is a key design objective for many AI researchers, it is common for people to view artificial intelligence as an autonomous system that does not rely on human participation. It is as though, if human intelligence is still participating in the system, then the AI isn't finished. Cybernetics may therefore offer a viewpoint for designing artificial intelligence in complex human systems where there is no desire to replace human intelligence with computational automation. This seemed especially apt in the context of supporting university administrators in supporting the wellbeing needs of their community during the COVID-19 pandemic.

It can be challenging to conceptualize the design of an AI system that makes such extensive use of human information processing and action. To conceptualize how an AI system can be designed in the context of a larger human system, we look to systems-thinking (Arnold & Wade, 2015, e.g.,) and cybernetic approaches (Krippendorff, 2007; von Foerster, 2003; Wiener, 1961, e.g.,). These perspectives point to how artificial systems may be designed to leverage human systems that are already functioning in a community, rather than trying to do everything autonomously. The cybernetic approach helps simplify the algorithmic design problem by focusing on a core process: generating a feedback loop between assessments of wellbeing and actions taken to enhance wellbeing. Furthermore, the cybernetics approach frees us from having to automate all processes into computational processes; we can design intelligence into a complex-sociotechnical system without having to make an entirely autonomous AI agent. Furthermore, we will show that such a system can be implemented rapidly and, over time, can be improved through iterative design, community feedback, and appropriate automation.

To summarize, based on theories of artificial intelligence and cybernetics, we sought to create an intelligent feedback loop capable of promoting community wellbeing. Figure 4.4 below visualizes the components that were involved in the feedback loops in our context. The design of the ability to

sense the state of the system will be discussed next after which the process of defining the *action space and processor* will be described.

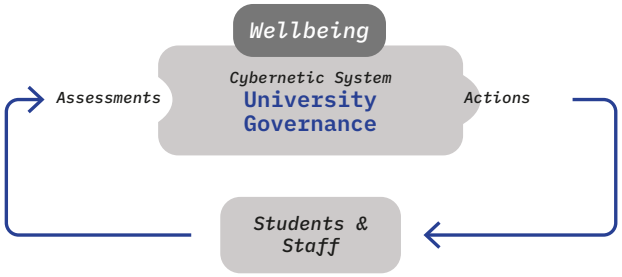


Figure 4.4: A schematic that visualizes the different components that were involved in the feedback loops in our context.

4.3. Case study: My Wellness Check

The COVID-19 pandemic has been a major shock to societies around the world, with far-reaching consequences for mental health and wellbeing. In particular, the COVID-19 pandemic appears to have measurably reduced rates of human wellbeing around the world (Brodeur, Clark, Fleche, & Powdthavee, 2021). The pandemic has also sparked interest in the role of large organizations in supporting wellbeing as an explicit criterion of organizational success. Universities, for instance, aim for a variety of organizational metrics of success (e.g., high graduation rates, large numbers of applicants, sustainability metrics, etc.). Increasingly, universities are recognizing student and staff wellbeing as explicit institutional priorities (Burns, Dagnall, & Holt, 2020; J. D. Lomas, Matzat, et al., 2021).

During the onset of the COVID-19 pandemic in early 2020, the university administration asked us how they could best support the wellbeing of their staff and students who were now locked inside their homes due to government restrictions. This led to a two-year, iterative, community-led design process of a system that helped support mental-health and wellbeing. In this section we will address parts of that process that may translate to other contexts. The generalizability lies in the parts that have emerged from our cybernetic process—such as the context-sensitive assessment—and translate conceptually, not factually.

The most essential requirement of our system was to create *the ability to sense wellbeing needs in the community.* This "sensor" would be at the heart of any future AI system or cybernetic system for community wellbeing.

Therefore, our primary goal was to design a wellbeing assessment instrument that was sensitive to the needs of our specific context—and capable of informing and motivating appropriate actions in response.

4.3.1. Theoretical approach to wellbeing

To develop our sensor, we did not choose one particular theory of human wellbeing (discussed in Box 6.5), but rather took a syncretic approach and drew from multiple theoretical traditions (e.g., Diener et al., 1999; Ryff & Singer, 2006; Seligman, 2011). This was justified because the goal of our assessment differed from the typical goal of conventional psychological approaches to measuring human wellbeing (discussed in Box 6.5), which is to create an accurate and theoretically valid measurement instrument. Instead, our goal was to create an actionable assessment: an assessment purposefully created to help inform and motivate concrete actions in the community to promote wellbeing.

Popular measures of wellbeing often focus on generalization. That is, they seek to validate a measure that can be used for comparing multiple contexts. Ed Diener's single item life satisfaction measure is a good example: on a scale of 0-10, how satisfied are you with your life as a whole? This measure (and its many variants) has been extremely useful for comparing wellbeing in different contexts. This measure is "actionable" insofar as a low score shows that something should be done. However, it does not give indications for specific actions. For this reason, we sought to devise new measures of wellbeing that were highly specific to the context of the community we sought to serve. We anticipated that a context-sensitive assessment would be more actionable (because it deals with specifics) as well as being more sensitive to the needs of the community.

Therefore, we used factorized models of wellbeing as an organizing principle to help identify concrete and specific questions that could support community action. Consider how various theoretical factors underpinning wellbeing may manifest within our context. For instance, many different models of human wellbeing recognize Material Wellbeing as an important factor of wellbeing (e.g., Sirgy, 2018). However, what material wellbeing means is likely to differ from one context to the next. In the context of wellbeing during COVID-19, for instance, we asked about the ergonomic quality of home workspaces—which can be seen as a causal indicator (C. S. Wong & Law, 1999)—as part of an effort to assess the influence of the home working environment on wellbeing. According to Mackenzie, Podsakoff, and Podsakoff (2011), it is specifically these sorts of causal indicators that belong in survey instruments.

Across all iterations, the My Wellness Check assessment considered a diverse range of indicators for community wellbeing: academic experience, anxiety, autonomy, behavior, belongingness, competence, coping strategies, COVID-19 measures, depression, drugs and alcohol, exercise, expected university support, finances, home working environment, life satisfaction, loneliness, mood, motivation, nutrition, optimism, overall physical health, personal growth, purposefulness, remote education, sleep, study performance and subjective mood.

Furthermore, we considered a wide range of existing surveys for the construction of assessment items were WEMWBS (Tennant et al., 2007), PERMA-profiler (Butler & Kern, 2016), SWLS (Diener et al., 1985, 1999), HILS (Kjell & Diener, 2021), WHO-5 (Topp, Østergaard, Søndergaard, & Bech, 2015), PWB (Ryff & Keyes, 1995; Ryff & Singer, 2006), WLS (Piliavin & Siegl, 2007), MIDUS (Radler & Ryff, 2010), and NSFH II (Springer & Hauser, 2006), CSSWQ (Renshaw, Long, & Cook, 2014), and Student WPQ (Williams, Pendlebury, Thomas, & Smith, 2017).

4.3.2. Community-led survey design

The design of our community wellbeing assessment combined traditional psychological methods for survey development (Boateng, Neilands, Frongillo, Melgar-Quiñonez, & Young, 2018) with a variety of human-centered design approaches, particularly community-led design methods (Costanza-Chock, 2020). By community-led design, we specifically mean that we involved community members and community leaders in the informal design and informal evaluation of the assessment instrument. Rather than approaching this in a strictly systematic manner (typical of psychological survey development), we encouraged various levels of university leadership to "weigh in" on the types of questions to be asked (typically in response to a proposed concrete example). Involving community leaders helped build "buy-in" and motivation for the deployment of the system.

To balance out the needs of community leaders with the needs of the community at large, we also put significant effort into gather diverse perspectives across the many iterations of the survey. At first, we used informal, semi-structured interviews with about 15 students and 7 staff members to gather perspectives on current needs and ideas regarding the academic experience and overall community wellbeing. These interviews were focused on identifying concrete and specific indicators associated with different theoretical factors of wellbeing. Together with the priorities of university leadership, these community interviews helped inform the focus of our initial set of survey questions. Once an initial survey experience was

developed, a sample of about 40 diverse students were asked to complete the entire survey over video chat. While offering informed consent and promising anonymity, we encouraged participants to comment aloud on individual survey items and give critical feedback. All content data was discarded for privacy reasons. However, this observational method was helpful for improving the relevance of the questions, reducing ambiguity about the meaning of questions and generally ensuring that important topics were not overlooked. Over the course of this development (which took approximately 4 weeks), many subtle iterations to the survey were made to ensure appropriate pacing and sequencing.

This iterative design and survey development continued even into the subsequent deployments of the survey, discussed in the following section. An example of the mobile user interface is presented in Figure 4.5; this shows the effort taken to create a motivating and positive survey experience. It also shows the tight integration of quantitative data collection procedures with opportunities to gather the voice of respondents in free text boxes.

After the initial deployments of the survey, statistical data about individual items made it possible to identify items that well predicted our central measure Life Satisfaction and items that did not. Furthermore, following the survey, all respondents were offered the opportunity to leave critical feedback about improving the survey. This strong focus on community-led iteration was hypothesized to create a better survey experience and to produce a more sensitive sensor of community wellbeing. As will be discussed, these hypotheses were tested through a controlled experiment.

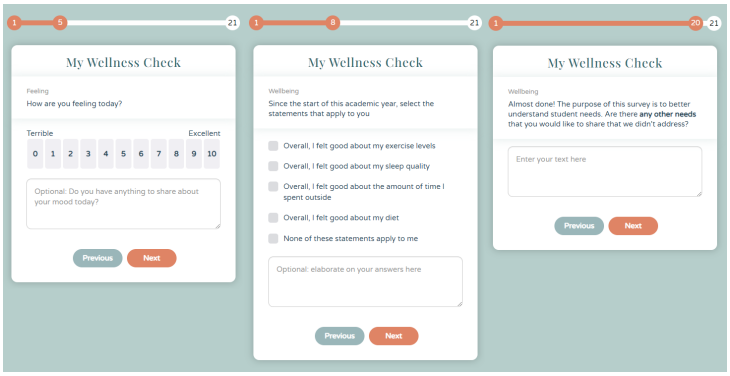


Figure 4.5: A selection of images depicting the appearance of the survey experience. The first screen shows a rating item about life satisfaction, the second, a checkbox item about physical wellbeing, and the third item shows a free text item about their additional wellbeing needs.

4.3.3. Deploying the assessment of community wellbeing

This section presents an overview of the deployment of our assessment of community wellbeing to 27,270 students and 6,347 staff members (in December 2021). Separate surveys were designed for students and staff. The number of participants and a summary of each iteration can be found in Table 1.

All students and staff received an email in both Dutch and English that invited them to participate in the study. The email contained a link that led them to an online version of the survey that could either be completed on a tablet, phone, or desktop. The welcome text of the assessment provided participants with information about the anonymity of their data (the limitations to guaranteeing their anonymity will be addressed in the discussion), the fact that the assessment was compliant with GDPR standards, and thus provided them with enough information to give their informed consent. All data were anonymized.

Table 4.2: An overview of data gathered in five iterations through scaled items about student wellbeing at Delft University of Technology. The range of each scale was from 0 ("Terrible") to 10 ("Excellent"), except for life satisfaction which was from 0 ("Very dissatisfied") to 10 ("Very satisfied")

Iteration	Date	n	Completion rate	# Q	# I
Staff 1	June 2020	2776	85% (2328)	24	56
Student 1	June 2020	3150	81% (2604)	25	79
Student 2	November 2020	3409	80% (2841)	26	82
Staff 2	December 2020	1826	89% (1622)	22	76
Student 3	March 2021	2877	77% (2221)	19	55
Staff 3	June 2021	2376	84% (2006)	25	49
Student 4	June 2021	2062	80% (1719)	19	79
Student 5	November 2021	1835	81% (1492)	19	91

Table 4.3

		Percent saying yes				
		Jun. 2020	Oct. 2020	Mar. 2021	Jun. 2021	Nov. 2021
Belongingness	I feel part of a community at TU Delft	44	28	20	24	26
	I often feel lonely	31	40	42	36	36
	I feel like I belong at TU Delft	57	41	41	38	33
	It often feels like no one at TU Delft cares about me	21	21	25	24	24
	I often feel like I don't have anyone to talk to				18	28
	I feel that my fellow students care about me and each other				39	27
	I have a good bond with one or more of my fellow students				67	60
	I would feel comfortable letting a professor know if I need help				26	
	Often, I felt like I could be myself around my fellow students					47
	Often, I felt left out					12
Overall Wellbeing	Overall. I felt good about my exercise levels	45	44	34	44	45
	Overall. I felt good about my sleep quality	52	51	48	46	46
	Overall. I felt good about my diet	61	62	54	53	53
	Overall. I often felt down	46	46	59	44	44
	I often worry too much	58	65	58	58	63
	Overall. I felt good about the amount of time I spent outside			26	43	40
	I feel like my stress levels are unsustainable				39	45
	Often, I felt relaxed					19
	Often, I didn't feel good about myself					32
Studies	I feel confident about graduating on time	50	45	42	42	35
	I am generally optimistic about the future	61	56	51	53	36
	I am happy with how I am performing in my studies	63	50	48	50	
	I am satisfied with my study/life balance	39	31	19	25	37
	I feel capable at what I do			35	39	
	I feel motivated to finish my current study program			57	58	58
	Overall, I felt I will be prepared to continue with my career successfully					34
	Overall, I felt satisfied with my online / offline balance					34

4.4. Qualitative results

In addition to statistical measures, we also designed the survey to support the collection of written text that represented the "voice" of students and staff. Open-ended questions were important for assessing community needs and also to gather possible actions that the university might take to help. These questions included: "What contributes to your sense of belonging and community at Delft University of Technology?;" "What aspects are you missing?" and; "Do you have ideas on how Delft University of Technology might help support student motivation?"

Open-ended questions were useful for eliciting concrete statements of student needs as well as serving as a source for specific ideas for the improvement of wellbeing. The quantitative data was useful for showing patterns across different university populations (e.g., could see the degree of physical health issues in international students v. local students). Written responses were often a source of specific ideas for organizational improvement. For instance: *"I really miss the in-between coffee chats with fellow students and the company. I want to see people"*; *"I really hope that hybrid learning can continue and that I can finish my master's degree while being in my home country."*; *"Organize silent discos with circles on the ground so that you can dance just in your own space!"; "I'm so happy to be back on campus, the facilities are awesome. However, it is hard to find a quiet place where I can talk in videocalls."* and*; Because of* [COVID-19], *some know each other very well and others don't. It's hard to join a group, especially if you don't know anyone at all. The university could help organize meeting groups for international students"*

4.5. Designing for community action

The previous section focused on the design of a sensor for community wellbeing. The purpose of this sensor was not just to generate a measure, but to serve in a cybernetic feedback loop that could motivate subsequent community actions—actions that could help contribute to improvements in community wellbeing.

Therefore, there were two core tasks required in designing for community action: 1) identify possible actions that could plausibly improve community wellbeing and; 2) motivate community actors to take appropriate actions. In practice, we made an effort to combine these two tasks together: when community actors were engaged in a process to help them identify useful actions for improving wellbeing, this was a key motivation for their subsequent actions.

Following each survey iteration, we held community-led design workshops with approximately 20 to 40 diverse online participants. Workshop participants included students elected to the student council, staff counselors (including psychologists and employees involved in mental health coaching), deans, upper administrators, the co-rector, and various other students and staff all from across the university. Prior to each workshop, each participant was given several hundred written responses to review. with the instruction to identify unique needs and their ideas for how to help. At the workshop, small groups synthesized and discussed these lists of needs and ideas. Following a whole group discussion about the "doability" and "urgency" of different ideas, the lists of needs and ideas were compiled together for presentation to university upper administration. This approach to collaborative data analysis. The aim was to analyze qualitative data to inform the communities about the wellbeing needs emerging from the survey and to collect ideas for improvement. To maximize the potential for action, we took special care to involve administrative decision-makers in reading and reviewing survey responses. Below, Table 4.4 highlights some of the ideas and the institutions that may be able to act in accordance with them.

Motivating community action also occurred through the presentation of data to various stakeholders. For instance, following data collection events, data presentations were made to the executive board of the university and to the board of education. Several policies can be directly linked to the results of our analyses. For instance, the university organized a 'Wellbeing Week' with various activities related to the outcomes of the report (i.e., supporting sleep, exercise and socializing). More concretely, as we found that the home working environment was a strong predictor of wellbeing, the university funded a program to provide ergonomic chairs and desks. A subtler example came from the impact of many students expressing that they'd appreciate a more *human* communication approach—e.g., the dean sending out emails asking students how they were doing, in a very personal manner. This finding resulted in a set of official guidance on changing the tone of voice in administrative emails. Next to administrative changes, action was also taken from a community perspective. For example, many PhD students that started in times of corona expressed they missed the opportunity to meet people and have "spontaneous social contact." This inspired a program called 'PhD Speed Dating' where PhDs were assigned to a random person on zoom so they could chat and expand their social network.

Table 4.4: Sample Action Space: A selection of initiatives with potential actors and target needs based on 2020 survey data during campus closures, covering guidance, communication, workspace, social interaction, health, and finance.

Area	Idea for Action	Actor
Guidance	Organize collective day starts, cultivate a morning routine	University culture and sports center
	An effectiveness tracking tool to overview your work progress	Heads of Education & Student Affairs
	Group for simultaneously graduating students (SCRUM meetings)	Graduate mentors, teachers, program directors, Library, Heads of Education & Student Affairs,
	Motivate students to go outside (RSI prevention)	University culture and sports center, wellbeing taskforce
	More available counselors and psychologists	Career & Counseling Service, Heads of Education & Student Affairs, Student Council
	An online chat box for talking to student-psychologists	Communications dept., Career & Counseling Service, ICT
	Provide people buddy/study groups and guidelines on healthy routines	Academic counselors, Program directors, graduations progress at Education & Student Affairs
	A platform where students can share their tips and tricks for how to cope with the pandemic	Career & Counseling Service, Education & Student Affairs, Student Communications dept., Wellbeing taskforce
	Increase (online) contact hours with teachers and mentors	Program directors
Communi-cation	Positive communication from departments, professors, teaching staff. Clear, regular, and motivational.	Communications dept., Student Communication dept., Department deans.
	COVID-19 website should be more up to date and accessible (including a weekly blog)	Communications dept., Wellbeing taskforce, and Career & Counseling Service
	Facilitate office hours and mentoring with digital tools like ‘Calendly’	Communications dept.„ Science center, X, Study and Student associations, and the Library
Workspace	More available, COVID-proof, working spaces on campus	Library, Heads of Education & Student Affair, Faculty deans, Facility Management, Faculty Secretary, Student Wellbeing Taskforce, and Alumni Office
	Support for home offices (with Wi-Fi etc.)	Process Manage, Faculties, Design Graduate projects
Social	Improved Peer Mentoring	Bachelor and Master coordinators and Communications dept.

	Online platform to meet other students (particularly for internationals)	Teachers and teaching assistants, Career & Counseling Service, Education & Student Affairs, Student Communications dept. and academic counselors
	Support for Community Organizers	Communications dept., Science Center, University culture and sports center, student and study associations and the library
	Online (drop-in) groups for activities (e.g., fitness at home)	University culture and sports center, Student Initiatives
	Online dinners and coffee moments	University culture and sports center and student and study associations
Health	45-minute zoom meetings, not an hour	Teachers, schedulers, and teaching assistants
	Sport and culture courses online	University culture and sports center and the Executive Board
	Educational activities that can be done without a computer	Teachers and teaching assistants
Financial	Share models for how to deal with student loans in the future	Student deans and counselors
	Workshops on CV creation and jobfinding	Career & Counseling Service
	Promote jobs as student assistants while other opportunities are shutdown	Student council, Student Recruiting Services, Communications dept.
	Make financial advisors available/provide financial survival tips	Student dean

Beyond these top-down policies, data were also used to motivate bottom-up community responses. Infographics were designed (see Figure 4.6) to communicate results to educators, staff and students at the university.

These materials didn't just incorporate quantitative survey data, but also the qualitative "voice" of students. Educators were invited to take these results in consideration when designing courses, lectures, and interactions with students. One responded: *"When something resonates with me and I empathize with it, I feel the urgency to act and implement improvements in my practice."*

Figure 4.6: Infographic excerpt shared with university educators, highlighting key qualitative and quantitative study findings on educational environment and wellbeing, distributed via a university newsletter with an explanatory summary.

When measures against COVID-19 allowed it, community workshops were also organized in person. A community of researchers and designers were engaged in a workshop inspired by the *World Café* format, which is based on the belief that people within an organization, if put in a social environment open to dialogue and exchange, can find solutions to even complicated issues (Löhr, Weinhardt, & Sieber, 2020).

We promoted and designed initiatives enabling the student community to take action and create impact on itself as well. All answers to the question "What daily routines are working well for you?" were collected and analyzed. This resulted in several visuals, which were shared in episodes once a week by study associations in their social media accounts, see Figure 4.7. The four episodes covered important student topics that emerged from the survey responses' analysis itself and consisted in first-person sentences about positive routines. The goal was to inspire students with routines that worked for their peers, hence having a higher chance to work for them as well. Other bottom-up results include student projects focused on wellbeing.

One student, for instance, created a recommendation system to help students optimize their living situation on a budget and promoted this to thousands of students.

Figure 4.7: Selection from the 'It Works for Me!' wellbeing campaign, showcasing student-contributed routines for wellbeing, in partnership with local student associations.

4.5.1. Experimental evaluation

A quantitative evaluation of our overall system remains challenging—for instance, it would be largely infeasible to conduct a controlled experiment involving multiple communities. For this reason, we have sought to quantitatively evaluate parts of the system. In this next section, we share the results of a controlled experiment conducted to compare our wellbeing assessment to other wellbeing assessments that were developed using standard psychological methods (Boateng et al., 2018). Because our community-led design methods so actively engaged diverse stakeholders in our community, there is a possibility that it may have led to reduced measurement efficacy. We chose the Warwick-Edinburgh Mental Wellbeing Scale (WEMWBS) and the College Student Subjective Wellbeing Questionnaire (CSSWQ) because they are widely used wellbeing assessment instruments and suitable in the context of our survey intervention. That is, we could not expect students to complete a full Positive and Negative Assessment Scale (PANAS)—to point to another widely used assessment instrument. It would have been too burdensome and distinct from the survey experience they were used to after four iterations (i.e., My Wellness Check).

In comparison to existing and validated assessments, we aimed to test the following two hypotheses. We predicted that our context-sensitive assessment of wellbeing would achieve:

1. Improved prediction of life satisfaction (a core measure of human wellbeing);

2. Improved measures of user experience.

These hypotheses were tested by randomly assigning samples of the research population to answer one of three questionnaires (WEMWBS, CSSWQ, MWC) after which their evaluation of their experience had been compared. This controlled experiment was conducted during the fourth iteration of the student survey, June 2021, see Table 4.5.

Procedure.
2,062 student participants were randomly assigned to different versions of the questionnaire: 12.5 percent of all participants would receive the WEMWBS, 12.5 percent would receive CSSWQ and the remainder would receive My Wellness Check (75 percent). These proportions were chosen because we have conducted our study in a real-world setting, meaning that the objective of the survey had to remain true to the initial goal—gathering data on student and staff wellbeing during COVID-19 to inform institutional action. The

Table 4.5: An overview of the different iterations of wellbeing assessment conducted at Delft University of Technology. #Q refers to the number of questions included and #I refers to the number of items included. The iteration that is discussed in this methods section is highlighted (Student 4). We have also included the iteration that took place after that because it was considered relevant for our discussion, particularly with regards to the section about sharing data back to the community—regardless, there were no experimental conditions to that iteration

Iteration	Date	n	Completion rate	# Q	# I
Staff 1	June 2020	2776	85% (2328)	24	56
Student 1	June 2020	3150	81% (2604)	25	79
Student 2	November 2020	3409	80% (2841)	26	82
Staff 2	December 2020	1826	89% (1622)	22	76
Student 3	March 2021	2877	77% (2221)	19	55
Staff 3	June 2021	2376	84% (2006)	25	49
Student 4	June 2021	2062	80% (1719)	19	79
Student 5	November 2021	1835	81% (1492)	19	91

WEMWBS and CSSWQ were chosen because they are frequently used and validated measures of global and domain-specific wellbeing.

Prior to beginning the experimental questions, all participants answered a common question about their life satisfaction. Following the experimental questions, participants were asked to complete seven questions about their questionnaire experience based on Stocké and Langfeldt (2004) and Baumgartner, Ruettgers, Hasler, Sonderegger, and Sauer (2021).

4.5.2. Experimental results

Table 4.6 shows that our context-sensitive assessment improved the overall sensitivity of the assessment and enhanced the survey experience for participants. To calculate sensitivity, we used a regression model to predict individual Life Satisfaction scores using the responses to questions from the three surveys. My Wellness Check (MWC) produced a higher R^2 (a measure of predictive fit) than the WEMWBS or CSSWQ. Including all MWC items in the model produces an R^2 of 0.75 while restricting the model to only the checkbox items (not the 0-10 scale questions) still produced an R^2 of 0.53, exceeding the R^2 of a model with all items in the WEMWBS (R^2 = 0.51) and a model with all items in the CSSWQ (R^2=0.42). Then, to compare participant ratings of the survey experience, a MANOVA showed a significant positive difference (p<0.0001) between MWC and WEMWBS and CSSWQ across all items listed in Table 6. The sole exception was that MWC was significantly more exhausting (p<.0001), see Table 6. This shows that participants taking the MWC survey found the experience to be of significantly greater value, significantly more engaging, significantly more worthwhile and significantly more fun. All statistical tests were conducted using JMP 16.

Table 4.6: An overview of the experimental results during iteration four. This experiment was twofold. Firstly, the top row shows the correlation each model had with life satisfaction expressed by their effect size (R^2)—i.e., the degree to which they were able to predict life satisfaction. Secondly, the table shows a comparison between the questionnaire experience of My Wellness Check (MWC), Warwick-Edinburgh Mental Wellbeing Scales (WEMWBS), and College Student Subjective Wellbeing Questionnaire (CSSWQ). The range for the question about satisfaction was 0 ("Very dissatisfied") to 10 ("Very satisfied"). For the other questionnaire experience questions, the range was from 1 ("Totally disagree") to 5 ("Totally agree").

	MWC	WEMWBS	CSSWQ
Correlation with Life Satisfaction expressed by R2	0.75	0.51	0.42
How satisfied were you with this questionnaire?	6.9 (1.7)	6.2 (2.0)	5.9 (1.9)
This questionnaire was of high quality	3.8 (0.8)	3.3 (1.0)	3.1 (0.9)
Completing this questionnaire was of some value to me	3.3 (1.0)	2.9 (1.1)	2.6 (1.0)
Completing this questionnaire was engaging for me	3.2 (1.0)	2.8 (1.1)	2.7 (1.0)
Completing this questionnaire was exhausting	2.2 (1.0)	1.8 (0.9)	1.9 (1.0)
Completing this questionnaire was worthwhile	3.5 (0.9)	3.2 (1.0)	3.0 (0.9)
Completing this questionnaire was fun	2.9 (1.0)	2.7 (0.9)	2.6 (1.0)
Number of questions	17	16	14
Average completion time in minutes (SD)	7:51 (9:45)	5:47 (7:17)	5:22 (9:40)

4.6. Discussion

The aim of this article is to demonstrate an approach to designing systems for improving community wellbeing. Based on a proposed framework for designing AI systems, we highlight the value of cybernetic theory when designing intelligent systems that involve complex human communities—in the sense that AI theory helps us to understand that feedback loops are a key feature of complex systems, and that involving humans in the design of feedback loops is necessary to create intelligent systems.

Based on this theoretical background, we share a case study in which we design an intelligent feedback loop to promote university student and staff wellbeing during COVID-19. Our work focuses on the use of community-led and human-centered design activities to produce "sensors" of wellbeing (a context-sensitive wellbeing assessment), "actuators" of wellbeing (a space of action that can be taken by different stakeholders in our community), and "processors" of wellbeing (which enable the transformation of sensor data

into action). In our case study, we describe the longitudinal fluctuation of community wellbeing over two years of the COVID-19 pandemic and explain the range of actions taken in response. To evaluate our efforts, we also share the results of a controlled experiment which indicate that our wellbeing assessment has improved sensitivity to wellbeing and provides an improved user experience in comparison to other "off-the-shelf" wellbeing assessment instruments. Our work suggests that community-led and human-centered design methods can play an important role in the design of AI systems to support community wellbeing. Figure 10, below, is a general schematic of our framework. Note that these steps apply to any complex system be they predominantly artificial or humane.

The remainder of our discussion shares a vision for describing how our "AI for Wellbeing" approach might generalize to other complex systems, including online social media systems and national governments. We then discuss several important limitations to this approach. Finally, we reflect on the relative merits of "cybernetic thinking" in the design of systems that seek to integrate human and machine intelligence.

4.6.1. A generalized Vision for designing intelligent systems to support community wellbeing

The case study in this article is specific to the context of our own university during the COVID-19 pandemic. The approaches and methods may be generally applicable to other universities or organizations that seek to prioritize community wellbeing. Beyond this, our framework and methods show promise for guiding the design of wellbeing feedback loops within other complex sociotechnical systems. In other words, our approach is not necessarily a blueprint for the next "COVID-24" but rather a way to understand how systems can deal with novel or urgent phenomena that affect the global society at large.

For instance, approaches taken here may offer insights for the integration of human wellbeing into the optimization of contemporary social media platforms like Facebook. To provide context, the CEO of Meta said: "we feel a responsibility to make sure our services aren't just fun to use, but also good for people's wellbeing." (Zuckerberg, 2018) This statement introduced a new "wellbeing" metric called Meaningful Social Interactions (MSI). Three years later, however, the 'Facebook Files' (Hagey & Horwitz, 2021) showed that there are still many aspects of social media services that harm user wellbeing. Our work demonstrates a system design approach and community-led design methods for human wellbeing feedback loops that may be useful in the design

of social media services and other sociotechnical systems. For instance, Facebook's MSI metric could be refined and expanded with wellbeing data collected through the community-led design methods and system design approach described in this article.

Our work may also generalize to societal governance, in general. During the COVID-19 pandemic, wellbeing in Europe fell to its lowest level in 40 years (Allas, Chinn, Sjatil, & Zimmerman, 2020). Wellbeing is often not explicitly valued in discussions of economic growth and decline. However, Allas et al. (2020) proposed a model of the monetary value of wellbeing by considering how much additional income a person would need to receive in order to raise their wellbeing by a desired amount. With this model, McKinsey estimated that wellbeing losses during the COVID-19 pandemic cost more than three times as much as the economic losses (i.e., reduction in GDP).

Increasingly, national governments are shifting from a single-minded focus on economic growth and turning to a more integrated 'wellbeing economy' focus (Fioramonti et al., 2022). Since the country of Bhutan changed its constitution in 2008 (Ura et al., 2012) to focus on "Gross National Happiness," the idea of wellbeing-based governance has become an intense topic of research. The Organization for Economic Co-operation and Development (OECD) promotes and maintains a measure of country-wide happiness and wellbeing that is used for ranking and policy purposes (Mizobuchi, 2014). Clearly, there are moves to make citizen wellbeing a more explicit measure of government success.

Here, we wish to communicate a design vision for governance for wellbeing that is focused on the experience of citizens. What do we want it to *feel* like to have governments or even smaller organizations work to maximize the wellbeing of their people? After all, there are always risks that come from focusing too much on optimizing a single metric (Rambur, Vallett, Cohen, & Tarule, 2013; Stray, 2020; Thomas & Uminsky, 2020). We turn to metaphor to communicate our design vision (Hekkert & van Dijk, 2011). An AI-based optimization of human wellbeing may feel unnerving, as though we have put a machine in charge of running society. Instead, our vision for optimizing societal wellbeing aims to feel more like a deliberative democratic process. Perhaps governments could use systematic wellbeing assessments as a participatory ritual (akin to voting day) to make it easier to "listen to the voices of the people." Then, we envision that the collective review of citizen needs and wants could feel more like deliberative, "town hall" democracy: a messy, time-consuming but intensely social process of figuring out *what do people need?* and *what actions can be taken to help?* Our case study

shows the potential for using human-centered and community-led methods to optimize wellbeing in organizations large and small; the above design vision aims to communicate how "AI for Wellbeing" might be extended to "governance for wellbeing" in a humanistic manner.

4.6.2. Limitations

The goal of an "AI for Wellbeing" system is to improve human wellbeing. Designing such a system requires, foremost, measurements of human wellbeing. But it also requires the ability to take actions in response to measurements. In an ideal world, the actions taken in response to wellbeing assessments would be 1) observable, 2) theoretically grounded (or to have a known mechanism of action and some predicted effect), and 3) empirically evaluated. However, in the case of a university during the COVID-19 pandemic, these criteria were not met. It was very difficult to know precisely what actions community members took in response to the assessment data. Further, few actions had a clearly defined theoretical model of how they were likely to impact wellbeing. Finally, none of the actions taken were evaluated statistically. Indeed, even if some actions were evaluated, there is little to suggest that they would have had the same effect at another point in time. As a result, it was difficult to evaluate the efficacy of our overall system. In other words, whether the human-centered activities conducted were the best *best of all possible* actions is indeterminable nor was it verifiable whether our approach was the *optimal* approach.

While we cannot make causal claims about the benefits of a wellbeing feedback loop, it may be possible to observe the functioning of our system like a prototypical cybernetic system, a thermostat. In a thermostat, a heater will stay on until the temperature reaches a desired range. In our case, once wellbeing returned to a range deemed "normal," the system goal had been reached and the university was able to shift resources to "business as usual." The community motivation for promoting wellbeing is analogous to the heater in this analogy. When wellbeing fell below a certain level, the university community was motivated to take a wide variety of actions. When wellbeing rose above an acceptable level, the motivation to focus on wellbeing was diminished. Like a thermostat, My Wellness Check turned up the motivation for action while the assessed need was high and reduced the motivation when the assessed need was low.

Designing this in a real-world university setting involved more than creating technologies, developing surveys and executing human-centered design methods. It also required a messy, informal and unscientific political engagement by ourselves, as researchers and designers. This

engagement was essential for getting buy-in and participation from multiple university stakeholders. Yet, through the dozens of meetings necessary to implement this system, we were able to leverage the community expertise of, for instance, psychological counselors, student advisors, human resource personnel, student council members and administrative leaders. This process is vastly more involved than simply "keeping humans in the loop" within a technical system. This limitation (or feature) will be relevant to the design of other intelligent systems for improving wellbeing within large-complex social environments: messy, democratic political processes may be required in addition to software development and user interface design. This creates new opportunities and demands for the appropriate role of human-centered designers in large, complex socio-technical systems.

4.6.3. Design thinking, AI thinking, and cybernetic thinking

In this section, informed by our case study, we discuss our perspective on the design of AI systems applied within complex human systems. In the field of human-centered design, "design thinking" is a process for creative problem solving that is often used in the design industry (Pressman, 2019). In parallel, "AI thinking" can be described as a process for computational problem solving that is often used in the AI industry; in rough strokes, "if there is a problem, AI may be the solution." However, many real-world problems are far more complex than AI algorithms can handle, particularly when AI is conceived as a fully autonomous agent. These real-world problems might include emotional engagement with other human participants or negotiating values or ethics. As a result, "AI thinking" has the potential to result in negative outcomes when it focuses AI designers on the production of fully autonomous systems that replace human intelligence with computational intelligence. A narrowed focus on algorithmic competence can result in the design of disembodied AI systems that fail to respect or leverage existing human capabilities in real-world systems (Gillespie, 2014; Krippendorff, 2021; Pangaro, 2021). For instance, "AI thinking" has produced product offerings promising to use complex data to provide medical diagnoses or educational recommendations. These offerings often fail because human doctors understandably distrust a "black box" diagnosis, just as teachers tend to distrust a "black box" curriculum selection (London, 2019; D. Wang et al., 2020). Instead, systems work better when they are not designed to be fully autonomous, but rather designed to provide services that can couple with existing workflows in an "unremarkable" manner (Yang, Steinfeld, & Zimmerman, 2019).

"Human-centered AI" offers the opportunity to design AI systems designed to work in concert with humans, not just to replace them. Humans will remain better at understanding people and responding to their emotional needs for the foreseeable future. Complex ethical or value-laden decisions will continue to require human stakeholders to negotiate. AI design methods need to consider the limitations of AI systems and design AI systems capable of working in concert with humans and existing organizations—supporting humans in their decision-making, rather than seeking to replace them. If AI systems are to be used by humans, AI systems need to be designed to meet the needs of human users.

However, we suggest that there are important conceptual issues that emerge from trying to strictly distinguish between artificial intelligence and human intelligence. Many artificial processes, rules or algorithms exist within organizations that are designed to produce intelligent outcomes. Should the artificial design of intelligent human processes be considered artificial intelligence? If artificial intelligence is defined as "the artificial design of intelligent processes," then this means that artificial intelligence does not require computational algorithms. Instead, algorithms might be merely written down and executed by humans. For example, "mastery learning" is an educational method involving a simple algorithm: if students demonstrate mastery on a topic test, they can proceed to the next topic, otherwise they are to continue to learn and master the topic at hand (Bloom, 1973). Mastery learning can be supported by computers, but it can also be implemented as a non-computational cybernetic feedback loop (e.g., just with teachers and paper tests).

Artificial intelligence could be defined as any kind of intelligent information process that is artificially designed—whether the process uses silicon microprocessors. This would broaden the scope of artificial intelligence to include all kinds of governing systems, not just those that rely on advanced computers. Consider the example of an autopilot; in the context of a self-driving car or even in a modern airplane, autopilot is certainly classified as a type of artificial intelligence. However, the first autopilot for an airplane was a mechanical system and was invented in 1912. If we take AI to mean "intelligent process that is artificially designed," then the implication is that there is a great deal of AI that doesn't involve computers. This could have far-reaching implications for how we think about AI systems and their impact on society.

For the sake of convention, some may wish to adhere to a popular conception of artificial intelligence that might be described as "an autonomous algorithmic system that uses advanced computational techniques

to accomplish non-trivial goals in a manner that human intelligence." In this case, the artificial design of intelligent processes that do not meet this definition might be termed "intelligent system design," rather than "artificial intelligence." Importantly, even non-computationally focused work may still contribute to the *field* of artificial intelligence, particularly when it demonstrates the application of artificial intelligence theory and methods to the design of intelligent systems.

In comparison to Design Thinking or AI Thinking, "Cybernetic Thinking" describes the design of intelligent systems, where the intelligence in the system relies on sensor/actuator feedback loops (van der Maden et al., 2022). By focusing on the design of information feedback loops, where a system's performance is used to modify the system's behavior, Cybernetic Thinking can be applied to the design of any goal-driven system, whether it is computational or not. For example, cybernetic thinking might be used in the design of educational systems (as in the description of mastery learning, above) or in employee performance reviews, where indicators of employee performance are used to modify ongoing performance. Cybernetic thinking may be valuable because it focuses on the dynamics of whole systems (including humans and machines), rather than naively focusing on popular computational algorithms. In our case, we found cybernetic thinking to be invaluable in the design of feedback loops to promote community wellbeing.

4.7. Conclusions

Wellbeing is not just an individual concern, but a community and a societal concern. By designing a system to assess and support community wellbeing in the times of COVID-19, we have demonstrated how to systematically prioritize wellbeing as an explicit objective within large, complex social systems. Our work makes the following key contributions:

First, based on theories of artificial intelligence and cybernetics, we contribute an approach to designing feedback loops to support human wellbeing at a community scale. This approach is highly relevant to sociotechnical systems that have large numbers of individuals. In the context of our case study, we are working in a very large university with over 30,000 students and staff.

Second, we contribute a specific case study applied to the context of COVID-19. This case study provides practical examples of the application of our approach, such as the community-led design of online surveys to generate valuable feedback in the form of wellbeing data from our university community. This feedback is then fed back to the community in the

form of qualitative assessments of need and summary statistics providing visual representations of how wellbeing changed over time and across sub-communities (e.g., academic and non-academic staff).

Third, we contribute an approach for using human wellbeing data to inform sociotechnical system design. We use our wellbeing data to generate insights and recommendations for improvements to our university's COVID-19 response. For example, we provide action recommendations to particular stakeholders in the university. This is an important contribution as it provides a concrete example of how wellbeing data can be used to improve sociotechnical systems.

Finally, the result of adopting a cybernetic framework and using human-centered and community-led design methods, is the development of a novel *context-sensitive* wellbeing assessment. To evaluate our instrument, we conducted a controlled experiment: in comparison to other validated wellbeing assessment instruments, we found that our context-sensitive wellbeing assessment was more highly rated by participants and also demonstrated stronger predictive validity. We also present qualitative evidence showing that our assessment yields more "actionable" data for motivating institutional and community action.

In our approach to designing interactive systems to support wellbeing, we have shifted from a focus on individual user needs to designing for communities and institutions. We have also shifted our thinking from designing a static product to designing an intelligent product-service system—a system designed to operate as a cybernetic loop within a large and complex sociotechnical system. Finally, our mindset shifted as we accepted that we were not the experts leading the design so much as facilitators of a community-led design process. These shifts may be subtle, but they represented an enormous leap from our initial perspectives on applying HCI, design and AI methods to create tools to support wellbeing during COVID-19. Our argument is that any future work on aligning AI systems with values like wellbeing and democracy will benefit from a similar process as the one presented in this paper.

Ethics Statement

The studies involving human participants were reviewed and approved by Human Research Ethics Committee of TU Delft. The participants provided their written informed consent to participate in this study.

5

Developing and Evaluating a Method for Positive AI

How can we determine if emerging methods live up to their ambitious aspirations? As AI design processes profoundly impact society, ensuring techniques translate principles into practice becomes critical. This chapter provides a timely empirical investigation of the nascent Positive AI framework which aims to embed wellbeing into AI systems through human-centered techniques. However, converting worthy intentions into positive outcomes requires rigorous validation. Building on learnings from the longitudinal case study of designing a sociotechnical wellbeing tool, this chapter advances a more in-depth evaluation of the Positive AI technique. Using the structured 'chain of evidence' framework proposed by Cash, Daalhuizen, and Hekkert (2023), this assessment traces across multiple studies to demonstrate the method's efficacy in translating wellbeing aspirations into practical design. This comprehensive evaluation ensures that the Positive AI approach not only adheres to theoretical ideals but also stands up to scrutiny, setting a precedent for validating novel methods before widespread application. By empirically grounding innovations in participatory AI alignment, we can progress responsible innovation visions into reality.

This chapter is under review for publication at *Artificial Intelligence for Engineering Design, Analysis and Manufacturing* as an article titled "Developing and Evaluating a Design Method for Positive Artificial Intelligence."

In an era where artificial intelligence (AI) permeates every facet of our lives, the imperative to steer AI development towards enhancing human wellbeing has never been more critical. However, the development of such Positive AI poses substantial challenges due to the current lack of mature methods for addressing these complexities. This article presents and evaluates the Positive AI design method aimed at addressing this gap. The method provides a human-centered process for translating wellbeing aspirations into concrete interventions. First, we explain the method's key steps: (1) contextualizing, (2) operationalizing, (3) designing, and (4) implementing supported by (5) continuous measurement for iterative feedback cycles. We then present a multiple-case study where novice designers applied the method, revealing strengths and weaknesses related to efficacy and usability. Next, an expert evaluation study assessed the quality of the resulting concepts, rating them moderately high for feasibility, desirability, and plausibility of achieving intended wellbeing benefits. Together, these studies provide preliminary validation of the method's ability to improve AI design, while surfacing areas needing refinement like developing support for complex steps. We finally propose adaptions for future iterations of the method such as the inclusion of wellbeing-related heuristics, suggesting avenues for future work. This human-centered approach shows promise for realizing the vision of 'AI for Wellbeing' that does not just avoid harm, but actively promotes human flourishing.

5.1. Introduction

It's 3 a.m. and the familiar prompt flashes across my screen: "Are you still watching?" The question jolts me back to reality. A wave of regret washes over me as I look at the time, knowing the early morning ahead. Yet here I am, lost in another late-night binge session. I can't help but wonder—what if this time spent indulging my streaming habits could have somehow contributed to my wellbeing instead of harming it?

Picture a streaming service that goes beyond providing entertainment and helps cultivate meaningful social connections. Or imagine a dating app designed to foster more than superficial hook-ups—one that nurtures emotional intelligence and healthy relationship skills with every swipe. Such systems, driven by artificial intelligence (AI), may seem idealistic. However, as AI becomes increasingly integrated into society, the demand for systems that are socially beneficial and promote human flourishing grows ("Living in a brave new AI era", 2023; Tomašev et al., 2020).

The impact of a technology often originates from the values inherent in its design (K. Crawford, 2021; Klenk, 2021). As such, the values manifested

in an AI system are the result of deliberate choices made during its design process (Fokkinga et al., 2020; van de Poel, 2020). Recognizing this, there is an emerging opportunity for establishing consensus on methodologies that purposefully integrate these values into the design of AI-driven systems (Morley et al., 2020).

In this article, we investigate the development of *artificial intelligence* through the lens of *positive* design (Desmet & Pohlmeyer, 2013). Consequently, we focus on the design of AI-driven systems that promote wellbeing.[1] Scholars advocate for using wellbeing as a practical guidepost for beneficial AI development, as it offers an empirically grounded, outcome-focused approach rooted in people's lived experiences. (Musikanski et al., 2020; Schiff, Ayesh, et al., 2020; Shahriari & Shahriari, 2017; Stray, 2020). Specifically, wellbeing frameworks compile multidimensional metrics that translate abstract principles into measurable indicators grounded in social science.

How to do this remains an open question. Therefore, in this article we discuss the development and evaluation of a 'Positive AI Design Method' integrating insights from positive design (Desmet & Pohlmeyer, 2013), positive computing (Calvo & Peters, 2014), human-centered design (Boy, 2017; Giacomin, 2014; Norman, 2005), and cybernetics (Dobbe et al., 2021; Glanville, 2014; Martelaro & Ju, 2018; Sweeting, 2016). Efforts to integrate ethical values into AI design, such as Value-Sensitive Design (VSD), have been recognized for their potential to align AI systems with broader societal values (Umbrello & van de Poel, 2021). However, these approaches often fall short in providing robust mechanisms to verify if the intended values are genuinely realized in AI design outcomes (Sadek, Calvo, & Mougenot, 2023c). To effectively design AI for wellbeing, however, it is imperative to rigorously assess its real-world impact (Peters, Vold, Robinson, & Calvo, 2020). Building on existing efforts, we investigate how the assessment of AI's wellbeing impact may enhance design approaches, developing a method that proactively integrates wellbeing as a core objective of AI design.

This article is primarily aimed at designers seeking to deepen their engagement with the field of AI and AI practitioners, defined in a broad sense, who are interested in designing AI systems that promote wellbeing. Through the lens of positive design, we explore methodologies and frameworks that

[1] While conceptualizing wellbeing is one of the challenges this method seeks to address, it can broadly be understood as "experiences of pleasure and purpose over time" (Dolan, 2014, p. 39). However, we draw from the third wave of positive psychology, which means that the method is attuned to the complexities and varied contextual factors that shape wellbeing (T. Lomas et al., 2021).

can bridge the gap between AI technology and human-centered design, offering insights and practical guidance for these audiences. By adopting the cybernetic perspective, we centralize the assessment of wellbeing impact within the AI design process. Our core objective is to evaluate the credibility and robustness of the Positive AI design method. To achieve this, we will follow the framework proposed by Cash et al. (2023), presenting a 'chain of evidence' that supports our approach. In doing so we aim to answer the following four research questions:

1. How might we standardize a method for designing AI that actively supports wellbeing?
2. What are the strengths and weaknesses of the method in practical applications?
3. To what extent does the method yield successful design outcomes?
4. How can future iterations of the method enhance its credibility and robustness?

The remainder of the paper is structured as follows:

- **Background.** Discusses definitions of AI and the need for a human-centered AI design approach, highlighting gaps in current methodologies and addressing *RQ1.*
- **Design Method** Outlines how the method was developed and refined, as well as the key steps of the method, answering *RQ1.*
- **Multiple-case Study.** Presents case studies of novice designers using the method, showcasing its practical strengths and weaknesses and addressing *RQ2.*
- **Expert Evaluation Study.** Reports on an expert evaluation of concepts from the Positive AI Design Method, directly relating to *RQ3.*
- **Discussion and Future Directions.** Discusses limitations and proposes enhancements for the method, reflecting on its contribution to human-centered AI and future research directions, answering *RQ4.*.

5.2. Background

In this section, we explore a definition of AI, scoping it as a special type of sociotechnical system through the concept of cybernetics. We further identify the key challenges and opportunities for incorporating human wellbeing into AI systems, setting the stage for the development of the Positive AI Design Method.

5.2.1. Artificial Intelligence

The term "Artificial Intelligence" carries a breadth of meanings that have evolved alongside its advancements. The essence of AI, as pointed out by AI pioneer John McCarthy, morphs as its applications become ubiquitous in everyday technology (Vardi, 2012). At the center is the notion of 'intelligence' itself. Although definitions vary, they commonly highlight abilities in reasoning, problem-solving, and adapting to new challenges (Sternberg, 2003). In an effort to integrate these recurring themes, AI researchers Legg and Hutter (2007, p. 9) propose defining intelligence as "an agent's ability to achieve goals in a wide range of environments." This perspective suggests that intelligence, fundamentally, is about an entity's adaptability and its proficiency in navigating a spectrum of scenarios to achieve its goals.

The 'artificial' aspect of AI lies in its *deliberate design*, contrasting with biological intelligence that naturally occurs in living organisms (Gabriel, 2020). As such, AI research focuses on *building* intelligent agents that choose actions to maximize performance based on received inputs and inherent knowledge, where agents perceive their environment through sensors and act upon it through actuators (Russell & Norvig, 2022, pp. 54–58).

Building on this understanding of AI, it is essential to recognize that AI systems exist within a complex sociotechnical context. Dobbe et al. (2021) highlight the frequent discrepancy between the promised benefits of AI systems and their actual consequences, termed the "sociotechnical gap." This gap arises from the divergence between socially necessary outcomes and what AI can technically achieve. For example, while a recommender system may aim to provide valuable suggestions to users, in practice, it could inadvertently promote misinformation or polarization.

To address this challenge, various scholars have proposed understanding AI as a sociotechnical system that encompasses not only its technical capabilities but also its limitations and the governance structures surrounding it (Dean, Gilbert, Lambert, & Zick, 2021; Dobbe et al., 2021; Krippendorff, 2023; Selbst, Boyd, Friedler, Venkatasubramanian, & Vertesi, 2019; Stray, 2020; van der Maden et al., 2022; van de Poel, 2020; Vassilakopoulou, 2020).

For instance, ChatGPT should be considered not merely in terms of its underlying model but also its user interface, the company behind it, public perceptions, and the various use cases and purposes it serves. As such, some of these scholars advocate for adopting a cybernetic perspective, which emphasizes the importance of feedback and adaptation in managing the inherent complexities of sociotechnical systems, giving rise to the inclusion of non-technical and natural entities (Dobbe et al., 2021; Krippendorff, 2023; Pangaro, 2021; van der Maden et al., 2022).

5.2.2. Cybernetics: AI as sociotechnical system

Cybernetics, which emerged in the 1940s, is an transdisciplinary field focused on communication, control, and circular causality in systems (Mindell, 2000). The term "cybernetics" is derived from the Greek infinitive "kybernao," meaning "to steer, navigate, or govern." A core concept in cybernetics is the feedback loop, which creates circular causality between a system's past, present, and future states (Wiener, 1961). At its core, cybernetics presents an alternative perspective to traditional AI design by emphasizing the symbiotic relationship between humans and machines within complex sociotechnical systems (M. Mead, 1968; von Foerster, 2003). It focuses on the dynamics of feedback loops, communication, and control mechanisms that underpin both biological and mechanical systems, proposing that understanding these can enhance the design and function of AI (Beer, Chiel, & Sterling, 1990; Sato, 1991).

By viewing AI through a cybernetic lens, designers are encouraged to consider AI not just as isolated algorithms but as part of an interconnected web of social, technological, and environmental factors (Dobbe et al., 2021; B. Scott, 2004). This perspective underscores the importance of adaptability, self-regulation, and the role of AI within broader societal systems, offering a holistic framework for understanding the challenges around designing AI. As such, the design of Positive AI interventions can go beyond the algorithm and even the platform itself. For example, companies may use the method to make adaptations to the recommender systems that govern their platforms, while smaller design firms may develop a third-party add-on that alters the interaction with a platform to support wellbeing. Interventions can even take place in the broader ecosystem, such as the development of institutional guidelines for use of ChatGPT in education—thus lowering student anxiety in using these tools while bolstering its potential educational impact.

5.2.3. Challenges of designing AI

When discussing the design AI systems, what specific process are we referring to? The development of AI is often conceptualized as a multi-stage life cycle, traditionally segmented into seven key stages,[2] with the design phase being integral in translating business cases into engineering requirements (Morley et al., 2020). This phase is critical, as it lays the groundwork for how AI systems will function and interact within human contexts. Norman and Stappers (2015) advocate for the involvement of designers throughout the entire development process of sociotechnical systems. However, this article primarily addresses the design phase. This emphasis does not detract from the importance of a holistic approach—which we strongly support—but rather aims to provide a detailed examination of the unique challenges and opportunities within this specific phase. By concentrating on the design phase, we aim to delve into the nuances that shape the early and critical decisions in AI development, understanding that these decisions have far-reaching implications for all subsequent stages.

Then, to return to the question at hand, designing effective AI systems poses unique challenges compared to traditional software systems. Fundamentally, the uncertainty surrounding AI capabilities and complexity of possible outputs makes it difficult to ideate, prototype, and evaluate human-AI interaction using standard HCI methods. As they point out, AI systems continue adapting after deployment, so designers struggle to anticipate changing behaviors across contexts. Additionally, the near-infinite output possibilities, especially for adaptive AI, mean traditional prototyping fails to capture the full range of behaviors and experiences (Yang et al., 2020).

Furthermore, as we will address later, effectively incorporating wellbeing into AI design demands engaging with user communities. However, integrating user communities into the AI development process is challenging because of technical complexities, the unpredictable evolution of AI technologies (Sadek et al., 2023b), significant communication gaps (Piorkowski et al., 2021), and lack of relevant expertise (Hsu et al., 2022). These difficulties are exacerbated when designing for values such as wellbeing, as they are complex, multifaceted (Schwartz et al., 2012), and interpreted differently across individuals (J. Graham, Haidt, & Nosek, 2009) and cultures (Sachdeva, Singh, & Medin, 2011). As such, there are many possible interpretations of values like fairness, trustworthiness, and empathy, as well as disagreement over their relative importance (Jakesch, Buçinca, Amershi, & Olteanu, 2022).

[2] A complete AI development life cycle includes seven stages: business and use-case development, design phase, data procurement, building, testing, deployment, and monitoring (Morley et al., 2020).

While there is broad aspiration towards high-level AI ethics principles like fairness and transparency, translating these into practice remains challenging (Morley, Kinsey, et al., 2021; Schiff, Rakova, Ayesh, Fanti, & Lennon, 2021). For example, a review of guidelines on AI ethics found extensive discussion of principles like transparency and fairness, but very little on technical explanations for achieving them (Hagendorff, 2020) Similarly, Schiff, Borenstein, Biddle, and Laas (2021) underscore the complexity of applying ethical principles like fairness and transparency across sectors, highlighting a gap in consensus on practical implementation, which directly impacts the integration of values such as wellbeing into AI systems. Bridging this divide between principles and practice remains an open research challenge. It requires developing methods that reduce the indeterminacy of abstract norms while retaining adaptability to diverse contexts (Jacobs & Huldtgren, 2021).

In this regard, we may look to VSD as a promising methodology for embedding abstract values such as privacy into concrete design specifications, thereby guiding AI systems to better serve and reflect the diverse needs of stakeholders while promoting inclusivity and human-centricity in technology (H. Zhu, Yu, Halfaker, & Terveen, 2018). For instance, Umbrello and van de Poel (2021) present a case-study in which they successfully translated crucial values like non-maleficence into actionable design criteria for a novel AI system.

However, Sadek et al. (2023b) note that in current VSD practices is their inability to effectively assess whether these values are genuinely reflected in the outcomes of AI systems, highlighting a significant shortfall in impact assessment mechanisms (both qualitatively as well as quantitatively). As we will later discuss, it is this gap that our method tries to fill for two reasons. First, for any impact-centered method (which arguably any value-oriented design project is), it is essential to establish causal links between interventions and system fluctuations (Fokkinga et al., 2020)—mere good intentions do not cut it. Second, as Schiff, Ayesh, et al. (2020) point out, impact measurement leads to evidence-based decision-making and promotes accountability, thus fostering iterative improvement. Now that we have an overview of the challenges related to designing AI, let us turn our attention to the additional challenges introduced by a focus on wellbeing.

5.2.4. Challenges of designing AI for Wellbeing

Designing AI systems specifically to enhance human wellbeing introduces additional complexities. That is, wellbeing is inherently multifaceted, variable across individuals, and manifests differently across cultural contexts,

making it difficult to define and design for in a measurable way (Halleröd & Seldén, 2013; Huppert, 2017). AI systems often optimize narrow objectives, making it hard to ensure they improve wellbeing holistically. Rather than promoting human flourishing broadly, they target limited metrics. This makes it one of the six grand challenges for human-centered AI (Ozmen Garibay et al., 2023). To address this challenges further, a recent article identified seven key challenges for designing AI for wellbeing (van der Maden, Lomas, Sadek, & Hekkert, 2024). They used cybernetics to group them into four categories, listed below and mapped to a simple schematic in Figure 5.1.

- **Conceptualization of wellbeing:** the challenges around choosing the appropriate theoretical paradigm for conceptualizing wellbeing and modeling wellbeing contextually given its complexity and unclear relationships between system components and wellbeing facets;
- **Operationalization of wellbeing:** the challenges of measuring wellbeing contextually with adaptive instruments, translating qualitative wellbeing data collected through community engagement to large-scale metrics suitable for optimization algorithms, correlating self-report and behavioral data collection, and reconciling the different paces at which wellbeing changes versus AI optimization occur;
- **Optimizing for wellbeing:** the challenges of making trade-offs between competing objectives (e.g., individual versus communal wellbeing) when optimizing AI systems for wellbeing, and dealing with the fundamental constraint that wellbeing changes slowly while AI optimization is rapid;
- **Designing AI actions that promote wellbeing:** the lack of mature methods and examples for putting wellbeing at the core of AI system design, beyond just avoiding harm. Most tools focus on alignment but lack concrete guidance on promoting human flourishing.

Several frameworks have been developed to guide the design of wellbeing-supportive technology, such as the framework by Wiese, Pohlmeyer, and Hekkert (2020) that maps wellbeing-enhancing activities (Lyubomirsky & Layous, 2013) to digital technology design, and the 'METUX' framework by Peters, Calvo, and Ryan (2018) that supports wellbeing in digital experiences. However, these methods were not specifically developed for AI and do not directly address all four challenges mentioned earlier. The IEEE-7010 standard (Schiff, Ayesh, et al., 2020) however, provides a more

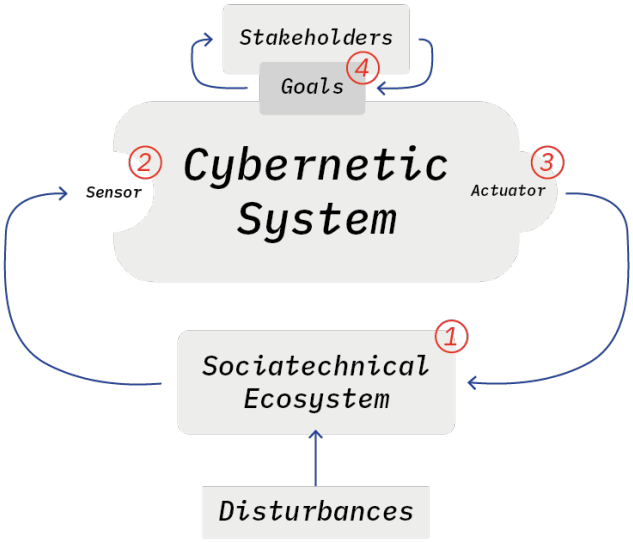


Figure 5.1: Shows a schematic representation of a cybernetic system. The different categories of challenges can be mapped onto this framework: (1) understanding the system context which entails modeling the relation between wellbeing of the systems constituents and its various components; (2) operationalizing said model of wellbeing; (3) designing interventions to actively promote operationalized model of wellbeing; and (4) retaining alignment with the overall goal. The latter refers to both challenges of algorithmic optimization as well as scrutinizing the objective (e.g., is the wellbeing objective still aligned to needs and desires of all relevant stakeholders?) Used with permission from van der Maden et al. (2024).

comprehensive approach tailored for the development and assessment of autonomous and intelligent systems (A/IS) with human wellbeing at the forefront. This standard offers an iterative Wellbeing Impact Assessment (WIA) process, stakeholder engagement, and a focus on wellbeing indicators across various domains. Through these facets, the standard addresses three of the four challenges by supporting the conceptualization, operationalization, and optimization of wellbeing. However, as a standard rather than a design method, it does not directly guide designers in translating insights

from the assessment into concrete design interventions.

Thus, our goal is to develop a method that not only builds upon existing frameworks, such as those outlined in the IEEE-7010 standard, but also harmonizes with recognized design and innovation approaches, including Design Thinking (Cross, 2023; Dorst, 2011) and the Double Diamond model (Council, 2007). By doing so, we hope to invite human-centered designers to the field of AI and bring human-centered design principles to the development of AI systems.

5.3. A Design Method for Positive AI

The Positive AI method is intended to provide designers with a structured process for developing AI systems that actively promote human wellbeing. It aims to address key challenges in conceptualizing, measuring, and designing wellbeing-supportive functionality into AI. It focuses on AI systems that people interact with daily, including curatorial AI (e.g., recommender systems), generative AI (e.g., ChatGPT), voice assistants (e.g., Alexa), among others. It is important to note that not all existing systems fail to address wellbeing, either by accident or on purpose; for example, Facebook and Google have made efforts to support wellbeing in their platforms, as discussed by Stray (2020). However, the Positive AI method aims to support any future endeavor where the pursuing wellbeing is an active goal not an afterthought, whether in existing or new platforms[3] such as Hume's Empathic Voice Interface (EVI).[4] Furthermore, the method does not cover autonomous vehicles, robots, or surveillance systems, as their embodiment presents challenges beyond our research scope.

By empowering designers and AI practitioners with concrete techniques, the method aims to create AI that measurably improves human thriving. It represents an initial attempt to address the lack of practical guidance in existing AI ethics literature specifically regarding enhancing wellbeing (Morley et al., 2020; Schiff, Rakova, Ayesh, Fanti, & Lennon, 2020).

5.3.1. Development of the method

The Positive AI method was developed using a cybernetic approach as an organizing framework following earlier discussions. Cybernetics views systems as cyclic processes of sensing states, comparing to goals, and taking action

[3] Calvo and Peters (2014) distinguish between active and dedicated wellbeing integration, noting that active integration into existing platforms presents additional challenges, as wellbeing goals must compete with preexisting objectives, such as those related to revenue.

[4] Available at https://www.hume.ai

(Mindell, 2000). This perspective enabled organizing the design challenges into distinct phases, with each phase addressing a different category of challenges (Fig. 5.1), while acknowledging the inherent entanglement present in complex sociotechnical systems (Dobbe et al., 2021).

We developed this method following a research-through-design process that drew inspiration from existing frameworks such as the earlier mentioned IEEE-7010 standard (Schiff, Ayesh, et al., 2020). The development involved collaboration between designers, researchers, and students over multiple projects. An initial two-year project designing a cybernetic system for institutional wellbeing during COVID-19 informed the first version of the method (van der Maden et al., 2023), which incorporated elements of the IEEE-7010 process, such as stakeholder engagement and a focus on wellbeing indicators across various domains. This early version was then refined through iteratively scrutinizing the methods efficacy and community feedback focused on streamlining the method steps, comprehensibility, and relevance. These various versions were then tested in five design courses given at the master's level, where student teams designed AI systems aiming to support wellbeing. These findings were then consolidated to present the version that is evaluated in this article.

Further, the Positive AI method is intended to complement and enhance typical design processes. For example, it parallels the empathize, define, ideate, prototype, test, and implement phases of Design Thinking (Cross, 2023; Dorst, 2011), with a specific focus on wellbeing and AI. Furthermore, the phases of our method align with the convergence and divergence characteristic of the Double Diamond framework(Council, 2007), while also emphasizing the iterative process inherent in most design strategies and frameworks.

The initial contextualization and operationalization phases strongly influence the subsequent ideation, prototyping and testing steps, discussed next. By grounding the design process in a contextualized understanding of wellbeing and corresponding metrics, the later activities remain anchored to the core goal of enhancing human flourishing.

5.3.2. Phases of the Positive AI Method

The Positive AI method is intended to provide designers with a structured process for developing AI systems that actively promote human wellbeing. In short, the method involves ensuring that AI systems are *sensitive* to factors of human wellbeing and *enabled* to support them. The five phases should help the designer to understand wellbeing in context (phase 1), to make it measurable (phase 2), to design systems (inter)actions that promote

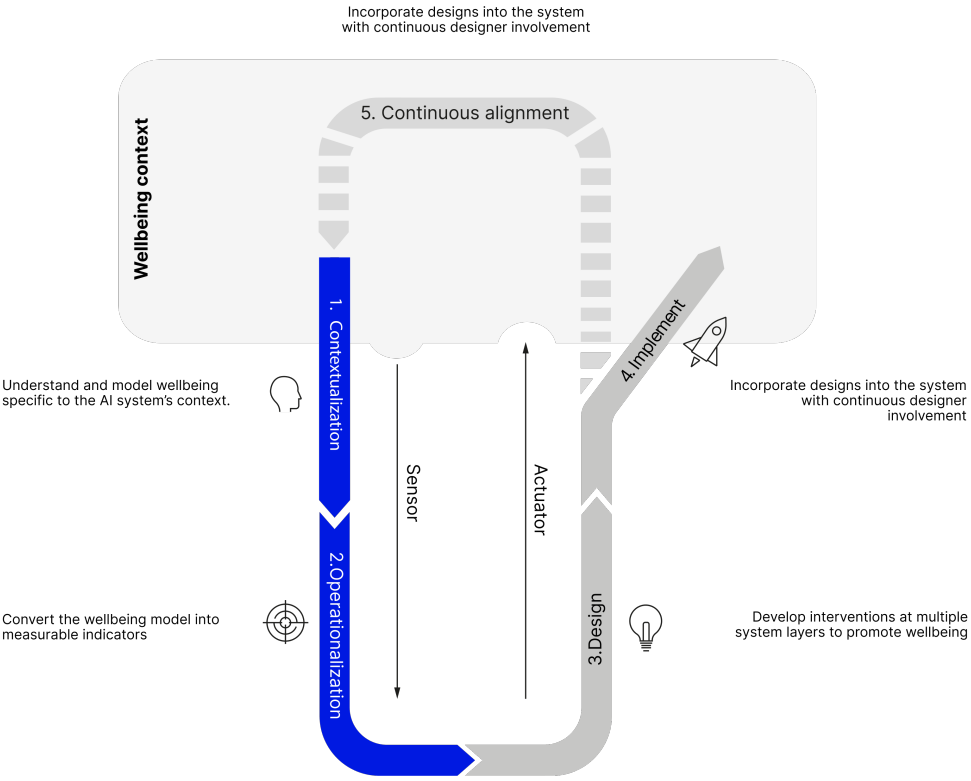


Figure 5.2: Diagram depicting the Positive AI method's cyclical approach, from contextualization to operationalization, design, implementation, and continuous alignment within the wellbeing context, illustrating the dynamic feedback loop between stages.

wellbeing (phase 3), to implement the designs (phase 4), and to sustain alignment (phase 5). Figure 5.2 shows an over of the methods phases with brief annotations of the content of each phase. A useful checklist with the activities and outcomes of the method can be found in Appendix.

Phase 1 - Contextualize: understanding wellbeing in context

To be able to sense wellbeing, we first need an understanding of what wellbeing is. However, wellbeing is complex, multifaceted, and manifests differently across contexts, making it difficult to conceptualize. For instance, despite there being overlap, how wellbeing manifests in an educational setting may differ from wellbeing in a healthcare environment. In education,

wellbeing may encompass a sense of purpose, self-efficacy, and belonging, while in healthcare it may manifest as physical health and effective pain management.

This complexity extends to AI systems and their broader contexts as well. Evidently, different AI systems, such as social media platforms or dating apps, influence wellbeing in different ways. For instance, the former might impact users' sense of social belonging and community engagement while the latter, might influence aspects of wellbeing related to relationships and self-esteem. In essence, wellbeing is shaped by the interaction between the user, their circumstances, and the specific AI system. Therefore, designers first need to understand how wellbeing manifests within their specific context—i.e., how the system they are (re)designing relates to the wellbeing of its user community.

A logical starting point is the extensive theoretical literature on wellbeing (e.g., Alexandrova, 2012, 2017; Cooke et al., 2016; Diener, 2019). This provides a wealth of information and can aid in the initial coupling of system components to wellbeing dimensions. However, the breadth of this literature can be overwhelming[5] and may not fully apply to the emergent nature of AI contexts (Kross et al., 2021; Stray et al., 2022). Consequently, designers must prioritize which aspects of wellbeing to focus on for their specific project.

To guide designers in prioritizing which aspects of wellbeing to focus on, we can follow the argument of S. Harris (2010) that they should focus first on the path that empirically contributes the least to suffering and the most to human flourishing.[6] This necessitates an empirical investigation to determine which dimensions of wellbeing are most relevant within the specific context—the low hanging fruit so to say.[7]

Before being able to engage in such an inquiry, it is essential to first

[5] I.e., due to wellbeing's complexity, there may be too many aspects and theories to consider in this stage of the design process.

[6] Harris' stance should not be interpreted as strictly utilitarian; he does not support maximizing happiness for the majority if it undermines the rights of minorities—a critical issue prevalent in today's AI landscape (K. Crawford, 2021). However, given the complexity of wellbeing, we must begin our inquiry with a practical choice: opting for paths that minimize suffering and enhance flourishing. There are other ethical frameworks to guide this process such as decolonial, feminist, or care ethics which would likely steer this inquiry in a different direction that warrants further exploration.

[7] This initial contextualization phase is arguably the "hardest" or at least the most "effortful," akin to the cold-start problem in recommender systems. Designers begin with minimal data, targeting broad, readily identifiable aspects of wellbeing and system components. Yet, with each iteration, their understanding deepens and responses become more nuanced, gradually enhancing the alignment between AI systems and wellbeing objectives.

develop an initial mapping of the AI system's components. This involves an analysis of the relevant elements within the system, such as its underlying algorithms, user interface design, data processing techniques, and output capabilities. This task is complex and unlikely to be fully resolved in the initial iteration. However, designers have employed a range of techniques to tackle this challenge, including stakeholder analysis Friedman, Kahn, and Borning (2009), context mapping Visser, Stappers, Van der Lugt, and Sanders (2005), and competitive analysis to gauge the effectiveness of various interventions Dalpiaz and Parente (2019). For example, in a social media platform, this would include examining how content algorithms shape user interactions and social norms, while for a dating app, it would involve analyzing the matchmaking algorithms and their impact on user experience and satisfaction. Each profession has a different understanding of what components of a system may yield which effects. Ideally, a Positive AI team would exist of a multidisciplinary group including human-centered designers which are trained to bring together diverse perspectives (Sadek et al., 2023a).

Then, having gained an understanding of the various components of our context, we can engage with the community to develop a more nuanced and detailed understanding of how those parts relate to specific facets of wellbeing. The goal here is to reveal which wellbeing facets seem most influenced or impacted within this particular environment. For this we can employ a slate of human-centered approaches such as interviews, focus groups, and observational studies to see which manifestations of wellbeing surface as most pronounced and how they relate to the components of the system. However, paraphrasing Alexandrova and Fabian (2022), how can we safeguard high scholarly standards of measurement while opening it up for lay participation? Therefore, we recommend grounding this investigation in rigorous methods such as proposed by Layard and De Neve (2023). By oscillating between contextually relevant indicators and scientifically established wellbeing metrics we can establish a constructive dialogue towards a nuanced yet scholarly model of wellbeing (Loveridge et al., 2020).

In other words, by thoroughly researching the user community and AI ecosystem from both practical and theoretical perspectives, designers can determine which components contribute most to wellbeing in that unique context. This allows designers to strategically focus design efforts on the wellbeing facets that are most relevant and impactful for that specific user community when conceptualizing and designing the AI system.

As a result, at the end of phase one (contextualization), designers will have established a contextual model of wellbeing which encompasses

hypothesized[8] relations between the most relevant facets of wellbeing and components of the context. This will be achieved by engaging in a conversation between the literature and the context. This dialectic channel can be opened through a plethora of human-centered approaches and is important in facilitating the continuous alignment which we will discuss in phase five. Before that, the next phase will address how to the abstract theoretical model measurable.

Phase 2 - Operationalize: making contextual wellbeing measurable

In this phase, the designer transforms the contextual model from an abstract concept to observable and actionable criteria. These criteria can then be assessed qualitatively (e.g., interviews, focus groups) and quantitatively (e.g., surveys, experiments). This process is invaluable, not only at the culmination of our design cycle, enabling the assessment of Positive AI interventions' impact, but also throughout the design process itself. It allows for ongoing evaluations of whether the prototypes are on track to achieve the desired outcomes and facilitates a continuous scrutiny of our contextual model of wellbeing.

More specifically, by developing observable criteria (both qualitative as well as quantitatively), we can refine our understanding of wellbeing, uncovering causal relationships and assess the effectiveness of design interventions. Through this operationalization process, we can empirically investigate the hypothesized connections of our contextual model defined in the previous phase. This mechanic of oscillating between theoretically defining a contextual model and empirically investigating it is core to psychological research into wellbeing (Diener & Michalos, 2009).

Note that it is important to recognize the distinction between measuring a phenomenon like wellbeing and its antecedents or determinants (Blijlevens et al., 2017). That is, we must separate measures of overall wellbeing from context-specific factors that influence it. For example, converting the abstract concept of "social connectedness" in the wellbeing model into a tangible metric might involve measuring both overall social wellbeing through validated scales, as well as specific indicators like the frequency and quality of an individual's social interactions. This allows us to capture the broad construct while also linking it to relevant contextual determinants, thereby revealing potential design spaces. It is in the combination of validated global

[8] "Hypothesized" in the sense that we will establish evidence for the causality over the duration of multiple cycles which allows us to assess the relation.

measures and context-relevant indicators that we find the actionable insights needed to understand and improve wellbeing (van der Maden et al., 2024).

Quantitative operationalizations of our contextual model are crucial for scaling our investigation and translating insights from local contexts to system-wide applications. For instance, when developing a wellbeing feature for a social media platform, initial tests with a user panel may not fully represent the broader community. A survey based on our operationalizations can validate the model's applicability at a system level. Additionally, integrating these operationalizations into AI system optimization processes, including algorithmic adjustments and managerial decisions, can significantly enhance system wellbeing. Operationalized metrics offer local indicators for system performance and wellbeing, facilitating their incorporation into optimization processes for more effective observation and refinement (Stray, 2020).

Finally, this process allows us to assess whether our design interventions produce their intended positive impacts on wellbeing. Such assessments can both be qualitative (e.g., observational studies—does the user engage with our interventions as intended) as well as quantitative (e.g., a controlled experiment comparing the wellbeing scores of two groups over time). This assessment process is essential for complex, interconnected design projects where various elements mutually influence each other (Fokkinga et al., 2020). By introducing interventions in a slow, incremental way, designers are able to couple wellbeing fluctuations to specific system components, hence grounding the Positive AI design process in empirical data.

Phase 3 - Designing: ideating and prototyping

With the contextual model and operationalized wellbeing metrics in hand, designers now have an idea of where in the system they can intervene to achieve specific wellbeing effects. Consider investigating how to intervene to improve the impact of ChatGPT, If observation studies and interviews reveal anxiety in "correct" usage of the tool in fear of being called a fraud, and feel a lack of authorship over what is produced, designers may look to ideate ideas to promote user empowerment and authenticity.

Nonetheless, choosing the right design direction can be challenging. To address this, 'scaling up the conversation' becomes crucial by verifying the hypothesized relationships identified in phase one at a system level. Employing quantitative methods through the operationalizations from phase two, such as user surveys, behavior tracking, and crowdsourced wellbeing ratings, can help pinpoint areas where interventions may yield the most impact. This approach not only identifies key focus areas but may also

generate new ideas from the target audience, as demonstrated in (van der Maden et al., 2023). Should this process not highlight impactful directions, revisiting earlier phases is advisable. It should also be emphasized that adopting a sociotechnical perspective on AI design allows interventions to occur across at least three distinct levels:

- **Experience design** - Crafting the overall user experience arc to positively affect wellbeing trajectories, either with interventions inside or outside the platform (e.g., guidelines for positive use of ChatGPT in education);
- **Interface design** – Leveraging the user interface for wellbeing-promoting interactions (e.g., "You're all caught up.");
- **Algorithm design** - Optimizing machine learning and recommendation algorithms to align with wellbeing facets influenced by the system.

There is no universally optimal design approach for this phase; the choice of technique is influenced by the specific context and scope of the design project. Designers and firms often have preferred methods, and are welcome to use these. However, to effectively kickstart the ideation process, particularly in tackling the previously discussed challenges of designing AI, we specifically recommend two resources. The 'AI meets Design Toolkit' (Piet & MOBGEN, 2019)[9] and the 'AI Design Kit' (Yildirim, Oh, et al., 2023)[10] stand out for their inclusion of generative prompts designed to aid in conceptualizing machine intelligence features. These tools are instrumental in facilitating a creative and informed ideation process.

At the end of this phase, designers will have produced a range of design strategies and artifacts that translate the operationalized model into actionable interventions aligned with wellbeing goals. The designer may end up with artifacts such as journey maps delineating goal-oriented user flows, wireframes illustrating proposed interfaces, interactive low fidelity prototypes, and explicit design principles encoding wellbeing aims. With these artifacts in hand, the designer then clearly communicates the guiding wellbeing goals and specific envisioned interactions to engineering teams for implementation. Ultimately, the success of this design phase lies in its ability to translate the operationalized model into a resonant yet actionable vision for design interventions that promote wellbeing.

[9] Available at https://aixdesign.co/toolkit
[10] Available at https://aixdesign.gumroad.com/l/toolkit

Phase 4 - Implement: integrating and testing interventions

In this implementation phase, the focus shifts to realizing the conceptualized interventions. This means further developing prototypes and testing them with users, thus putting the designers vision in effect. In the design of sociotechnical systems, it is important that designers are included in the implementation phase (Norman & Stappers, 2015; Sadek et al., 2023b). It is not solely the domain of development and engineering teams to bring these designs to life; designers must maintain a hands-on presence to guide and refine the implementation.

Specifically, the design artifacts and principles produced in phase 3 provide critical guidance during the implementation phase. In a collaborative effort with these designers, engineers may utilize these tangible visions of the system's form and function to construct the necessary components ready for user testing. Additionally, designers refer to these artifacts to steer the ongoing development, ensuring alignment with the wellbeing-centric principles encoded within them. For example, by comparing implemented features with the prototypes and design criteria, designers can identify divergence from the intended optimized interactions. This ability to reference the codified vision facilitates course-correcting implementations back into alignment.

By staying engaged through the implementation phase, designers are better positioned to address any unforeseen challenges that emerge. This proactive approach ensures that the wellbeing impacts, carefully planned in the design phase, are fully realized in the final product. By avoiding shortcuts or efficiency concessions, we safeguard the integrity of our project's goals. If these concessions were made, it would significantly undermine the very purpose of our endeavor. The sustained participation of designers is essential in bridging the gap between user needs, technical constraints, and the original design vision. In essence, the artifacts and guiding principles developed in phase 3 play a pivotal role in keeping the implementation firmly anchored to the wellbeing impacts, ensuring these principles are not lost but rather brought to life in the final integration.

Phase 5 - Reiterate: sustaining continuous alignment

Finally, maintaining alignment with the system's wellbeing context, as previously discussed, is crucial. This involves continually assessing whether our interventions meet their intended goals and if the wellbeing model remains applicable. Such evaluations allow us to stay attuned to changes in the

wellbeing context and uncover new opportunities for positive intervention. This process leverages established communication channels and operates on two distinct levels.

At the process level, designers should continually engage users and communities during contextualization and design activities. Human-centered methods like interviews, focus groups and co-design workshops enable aligning design decisions with community goals as they evolve.

At the system level, implementation marks the end of one iterative cycle. As the loop gets tighter through repeated iterations, the need for major interventions tends to diminish as positive adjustments accumulate. Nonetheless, the designer can step back, evaluate what occurred in relation to the wellbeing model, and determine needs for the next round. Does the contextual model require updating? Were key perspectives missing?

Restarting the loop enables revisiting the contextual understanding and community connections to realign priorities. By continuously iterating alignment at process and system levels, the approach maintains a pulse on emerging wellbeing impacts as user needs and technological capabilities shift. This cycling sustains the contextual accuracy and relevance of the wellbeing focus over time.

To effectively implement the Positive AI framework, it is crucial to consider the composition and collaboration of the team. Ideally, the team should consist of individuals from diverse disciplines, including designers, AI experts, domain specialists, and user representatives. Establishing clear communication channels and protocols is essential to facilitate effective collaboration among team members (Morley, Kinsey, et al., 2021; Sadek et al., 2023a). Regular meetings, workshops, and documentation can help bridge disciplinary gaps and ensure a shared understanding of project goals and progress (van Dijk & van der Lugt, 2013). By carefully considering team composition, communication, and stakeholder engagement, the Positive AI framework can be more effectively operationalized to address real-world challenges and opportunities.

5.3.3. Method applied to fictional example of a streaming platform

Imagining applying the Positive AI method to align a streaming platform with wellbeing, we would start by reviewing literature on video platforms and human-AI interaction to compile a list of key features and hypothesize their impacts on wellbeing. For example, studies suggest personalized video recommendations can sometimes limit users' openness to new perspectives and create filter bubbles. In contrast, features like custom video playlists may

boost users' feelings of autonomy and control over their viewing experiences (Möller, Trilling, Helberger, & van Es, 2020). So at this point, our initial theoretical model includes hypothesized relationships between features like recommendations and playlists and wellbeing aspects like openness and autonomy.

To refine this initial theoretical model we could conduct comprehensive user research through interviews, focus groups, and surveys. This research aims to gather first-hand accounts of how the platform's features influence wellbeing, focusing equally on identifying challenges and uncovering opportunities. By engaging a diverse sample of users, we aim to understand a wide range of perspectives and experiences.

We would then operationalize wellbeing by selecting validated scales like the Personal Growth Initiative Scale (Freitas et al., 2018) to quantify growth and openness to new ideas. Additionally, we would develop context-specific metrics, such as an aggregated playlist complexity score. This local metric could be calculated from factors such as breadth of topics, diversity of creators, and degree of organizational structure in users' video playlists. It would serve as an indicator of the level of perceived control over viewing experiences.

Equipped with this contextualized model of the platform's wellbeing impacts, we could propose targeted interventions to optimize the scales and metrics. For instance, one could suggest an algorithmic adjustment that sporadically introduces unexpected video recommendations, motivating users to explore content beyond their regular preferences, thereby potentially elevating personal growth metrics.

To implement such proposals, collaborative sessions with designers, engineers, and users would allow iteratively developing and refining features based on observed wellbeing impacts and user feedback. Designers would facilitate participatory design workshops to envision algorithm tweaks and interface changes. Engineers would build required components and monitor the system. Users would provide perspectives to ensure changes align with their values and goals.

By continuously revisiting the contextual model and indicators, the platform could incrementally adapt their AI systems to support multidimensional wellbeing objectives rooted in an adaptive, nuanced understanding of diverse user needs and values. The balanced set of global and local metrics would enable holistically tracking progress. A design cycle like this could be a first step of moving a platform beyond user satisfaction to a more complete alignment to wellbeing.

5.4. Multiple-case study

Three design students applied the Positive AI Design method for their master graduation projects at Delft University of Technology. They redesigned or build upon (parts) of existing ubiquitous AI systems to support wellbeing. These redesigns varied in the intervention level (i.e., from UI interventions to suggestions for changes in the algorithm) and consequently their impact on wellbeing. None of the students had experience with designing AI nor for wellbeing.

Student 1 chose to work in the context of dating apps and was specifically interested in how these could optimize for other components of human identity beyond looks. For example, dating apps have hidden mechanics that prioritize physical appearance in their matching algorithms (Klincewicz, Frank, & Jane, 2022; Parisi & Comunello, 2020). She wanted to explore how they could also factor in and foster other aspects of identity.

Student 2 chose the context of nutritional and food apps that, for example, track calorie-intake and suggest recipes. Such apps tend to prioritize nutritional intake as a proxy for wellbeing. However, a hyperfocus on nutritional intake may have negative effects on wellbeing while there are other aspects to eating that may actually benefit it that are currently often neglected (König, Attig, Franke, & Renner, 2021). Therefore, she aimed to broaden their scope to also account for the social and emotional aspects.

Student 3 chose the context of music streaming platforms. The AI recommendation engines in such platforms tend to provide the user with "more of the same" based on listening history and patterns (Tommasel, Rodriguez, & Godoy, 2022). However, music has powerful potential to influence one's personality, functioning, and understanding of the world. She hoped to leverage the existing AI in such a platform to encourage this kind of personal growth and exploration beyond repetitive patterns.

The goal of the multiple-case study is to assess two aspects of the design process itself. First, it examines the **efficacy** of the method, looking at whether the designers demonstrate thoughtful understanding of how their decisions potentially impact wellbeing. Specifically, does the method successfully elicit the desired focus on wellbeing considerations from designers, rather than other behaviors? Second, the study evaluates **usability**[11] aspects of the process, such as avoiding unnecessary detours or delays. This refers to whether designers understand the steps involved

[11] The term "usability" is sometimes used synonymously with "efficiency" in the literature. However, the concepts of efficiency, efficacy, and effectiveness are often conflated (Zidane & Olsson, 2017). To avoid confusion, this paper uses the term "usability" as it encompasses efficiency and has been used to refer to method efficiency by Cash et al. (2023)

and feel confident executing them. In other words, the efficacy assessment examines if the method shapes designer behaviors as intended, while the usability assessment looks at how easily and efficiently designers can apply the process.

5.4.1. Procedure

Three student projects utilizing the Positive AI method were initiated over a three week period. Each student received a personalized introduction to the project and was provided with reference materials including an overview document of the method, recommended literature, and the 'Positive Design Reference Guide' (Jaramillo et al., 2015) to support their research. When the third student commenced their project, a collaborative kick-off meeting was held to establish a shared understanding of the method, address questions, and align expectations across the projects.

Over the remainder of the project, the students met weekly with the supervisory team for guidance. This structured approach aimed to sufficiently equip the students with the necessary understanding and resources to effectively apply the Positive AI method within their individual projects. By providing one-on-one introductions, reference materials, a collaborative kick-off, and ongoing supervision, the aim was to support the students in comprehensively and successfully utilizing the Positive AI approach.

Then, to gather information on the method efficacy and usability, multiple sources of information were consulted. These included observations taken during the weekly meetings, progress reports and presentations, the final design outcomes and reports, and three recorded and transcribed one-on-one interviews with the students. The weekly meetings provided a platform for the students to share problems they encountered. Oftentimes, they faced similar hurdles, which were carefully documented. The progress reports and presentations served as useful post-hoc data to examine how the students were dealing with challenges, how the process developed, and how the designs developed over time. Finally, the interviews aimed to substantiate key themes identified throughout the project period. The first author analyzed the various data sources and shared the findings with the students to ensure accuracy. Before presenting these results, we provide a brief overview of the final design outcomes.

5.4.2. Materials: Design outcomes

In this section, we briefly summarize the final design concepts resulting from the three student projects applying the Positive AI method.[12] We present the core functionality and wellbeing goals addressed by each design to provide context before examining the process evaluation findings.

MiHue

Student 1 designed a dating app called "MiHue" that leverages AI to enhance users' experience of autonomy and relatedness. The core concept balances the needs for uniqueness (autonomy) and connection (relatedness) by highlighting individuality within similarity.

To identify which wellbeing aspects to focus on Student 1 began by conducting thorough research to understand users' wellbeing needs and experiences in using dating apps. Her literature analysis revealed autonomy and relatedness as salient wellbeing facets impacted by dating platforms. To further refine her contextual understanding of wellbeing, she also held multiple generative workshops with target users. During these co-design activities, participants also ideated improvements focused on supporting self-expression and social bonds.

Synthesizing her findings, Student 1 operationalized autonomy and relatedness within her context as the ability for users to express their unique attributes (autonomy) and to find meaningful connections based on shared interests or experiences (relatedness). She then formulated her design directions formulated a design direction aimed at enhancing social connection by highlighting individuality within similarity. This approach focused on promoting a shared connection through uniqueness and common ground, using the AIxD Ideation cards (Piet & MOBGEN, 2019) to link technology capabilities with desired wellbeing outcomes. This led to new features that encouraged users to share more personal and diverse aspects of their identities, beyond physical appearance. One key feature was an improved profile creation tool that prompted users to respond to creative and introspective (AI generated) questions, facilitating deeper self-expression. Another feature was an algorithm designed to match individuals not only based on mutual interests but also on shared values and life goals, aiming to foster more substantial and meaningful connections.

To closely mirror real-world application, she developed a strategy for implementing the novel features within either an existing app or as a standalone platform. Subsequently, she conducted user testing with an

[12] The showcases of the projects are available at https://t.ly/Fn8hn.

interactive prototype (designed using Figma). These tests included questions based her earlier operationalization of the contextual wellbeing model such as those related to autonomy (e.g., "How well does the app allow you to express your true self?") and relatedness ("Can you describe any interactions you had through the app that made you feel understood or belonged?"). This process was aimed at gathering feedback to refine the recommendations for subsequent iterations, thereby embodying the method's emphasis on continuous alignment. Her subsequent recommendations for the next phase emphasized expanding the focus to encompass additional aspects of wellbeing not initially covered but highlighted in the theoretical model, such as self-acceptance, positive emotions, and physical health. Moreover, she underscored the importance of including diverse user groups, particularly minorities, and considering gender differences, to ensure a more inclusive and comprehensive approach to enhancing wellbeing through the app's usage.

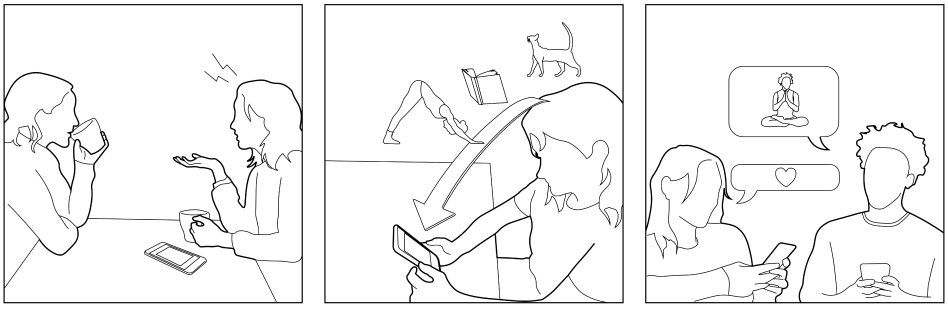


Figure 5.3: Three visuals used to illustrate key aspects of MiHue's journey as presented in the expert study: (a) The protagonist's frustration with current dating apps that focus on looks over personality; (b) The protagonist entering their interests during MiHue's enhanced account creation process that encourages authentic self-representation; (c) The protagonist matching with someone who shares common interests, as highlighted by MiHue's features that spotlight unique and shared traits between users to foster meaningful connections.

FoodVibe

Student 2 designed an app called "FoodVibe" that uses AI capabilities including facial recognition, natural language processing, and machine learning to provide personalized recipe and dining recommendations tailored

to users' specific social contexts and past preferences. The adaptive system aims to promote wellbeing through home dining by through mindful eating and nurturing social connections through shared meals. The idea for FoodVibe originated from the fact that existing nutritional and dieting apps tend to emphasize calorie intake and nutritional intake rather than other aspects of eating that affect wellbeing.

To identify what wellbeing aspects to prioritize in her design, Student 1 conducted a literature study as well as experience sampling (Van Berkel, Ferreira, & Kostakos, 2017). Her literature analysis revealed mindfulness, social connections, autonomy, and engagement as salient yet overlooked facets. She then combined this information with results from the sampling study to map a user journey specifically aimed at understanding when certain wellbeing experiences may occur. Next, to refine her understanding, she hosted two generative workshops where users emphasized the value of reflection, awareness, and social aspects around meals. This led her to operationalize wellbeing in her context (eating at home) as being present (e.g., engaging with your meal instead of the television) and having a sense of belonging (e.g., feeling related to your family when cooking a nostalgic dish).

This led to the design of FoodVibe which enhances wellbeing by encouraging mindful dining at home, giving users autonomy in their food choices, and deepening connections with dining companions. Utilizing AI, FoodVibe personalizes recipe suggestions by analyzing users' dietary preferences, the people they eat with (identified through facial recognition), and past meal satisfaction. Its features focus on personalizing meal recommendations, enriching social interactions by connecting with friends within the app and aligning meal choices with the group's tastes, and promoting self-reflection on dining experiences to boost wellbeing aspects like autonomy, positive relationships, and mindfulness.

The final design was evaluated through user testing with a high-fidelity prototype. The evaluation focused on whether the app achieved its design vision and goals, emphasizing the enhancement of eating experiences and perceived wellbeing of healthy-eating-app users. The effectiveness of FoodVibe was assessed based on metrics related to autonomy, positive relationships, mindfulness, engagement, fun, and the overall usability and desirability of the app. These metrics were derived from theoretical models and earlier phases of the research. This led to recommendations for subsequent iterations focusing on user experience improvements, like using avatars for privacy, broadening wellbeing theories to various dining contexts, boosting AI accuracy for tailored recommendations, and conducting thorough user testing for long-

term wellbeing impacts.

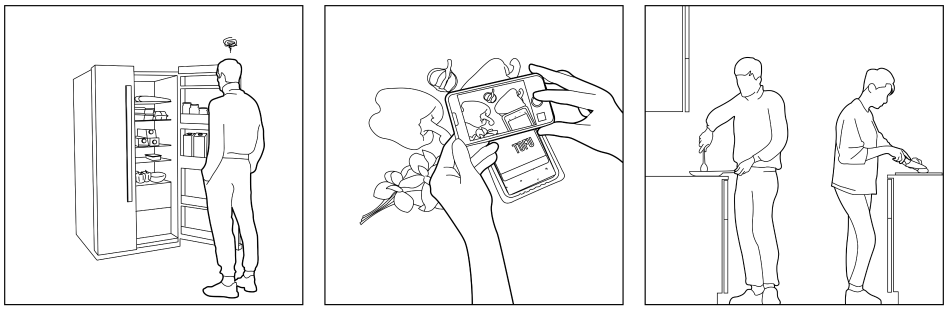


Figure 5.4: Three visuals used to illustrate key aspects of the FoodVibe journey as presented in the expert study (a) A user frustrated with nutritional limitations when deciding what to cook; (b) The user utilizing FoodVibe's 'Recipe Generator' by taking photos of ingredients on-hand so the app can suggest customized recipes; (c) Two people cooking together in the kitchen, representing FoodVibe's goal of promoting wellbeing through shared meals and human connections.

Explore More

Student 3 designed a new *Spotify* feature called "Explore More" that uses the platform's algorithms and extensive music catalog to guide listeners through unfamiliar genres in a personalized way. This idea stems from the observation that existing personalization mechanics, such as "Discover Weekly" playlists, tend to converge on a type of music which over time can get uninspiring. The goal is to expand users' musical tastes and perspectives to foster personal growth and empathy.

She began by analyzing the current landscape of music streaming service features and linked these to wellbeing literature. Her analysis revealed personal development, specifically through music's potential to facilitate self-discovery, as an impactful yet underutilized application for enhancing user wellbeing. Synthesizing her contextual findings, the student recognized limitations of Spotify's existing personalized discovery playlists driven by recommender systems, which can restrict users within narrow musical preferences over time.

Upon recognizing these limitations, she directed her efforts toward intentionally utilizing music's capacity for perspective expansion and self-

discovery to promote personal growth. In other words, she operationalized wellbeing as increased engagement with unfamiliar music genres, or in other words, she hypothesized that exposure diversity would lead to self-development. This inspired the design of Explore More, a feature that could either be integrated into a service like Spotify or as a stand-alone third-party interface. Features included an interactive genre map to visually navigate through unexplored musical territories, guided discovery paths offering sequences of new genres tailored to the user's tastes, personalized recommendations within those genres to ensure a resonant listening experience, self-reflection prompts aimed at deepening users' introspective engagement with the music, and a feedback and adaptation mechanism to refine future explorations based on users' experiences and preferences.

She developed an interactive Figma prototype for user testing, revealing key insights into navigation ease, the effectiveness of discovery paths, and the resonance of music recommendations. Self-reflection prompts were particularly noted for deepening users' personal insights and musical connections. Based on feedback, she recommended refining the UI for better navigation, enhancing the recommendation algorithm for tailored music exploration, and deepening self-reflection prompts for richer introspection. Additionally, she proposed adaptive feedback mechanisms to align the exploration journey with users' changing tastes, ensuring Explore More effectively supports personal growth and musical discovery.

5.4.3. Results of the case studies

To reiterate, a key aspect in evaluating the method is assessing its efficacy by examining whether the designer shows understanding of wellbeing impacts and thoughtful consideration of them in their design decisions, as well as grasping the relationships between wellbeing dimensions and system components. Whereas method usability refers to aspects like avoiding unnecessary detours or delays, understanding the steps involved, and feeling confident executing the method.

The multiple-case study revealed both strengths and weaknesses of the Positive AI Design method when applied by novice designers. In terms of method efficacy, students initially struggled to feel confident in their comprehension of the wellbeing literature. This was partly due to their unfamiliarity with the field. That is, the breadth of literature was overwhelming, causing uncertainty about when enough research had been done to proceed. This resulted in hesitancy during key stages as students were unsure they grasped concepts well enough to move forward.

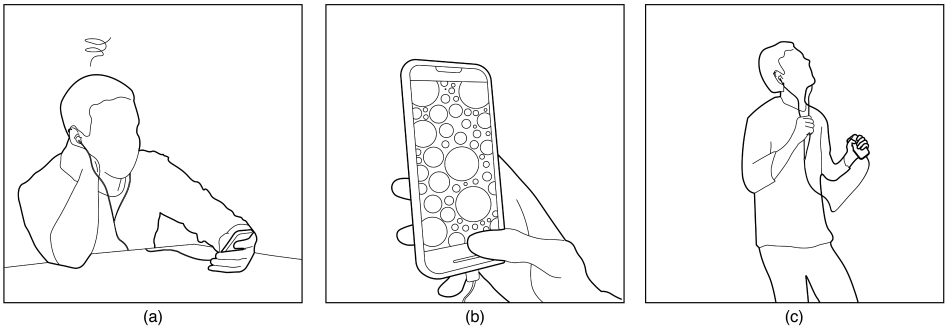


Figure 5.5: Three visuals used to illustrate key aspects of the Explore More journey as presented in the expert study: (a) A bored user unsure what to listen to; (b) An interactive map of music genres that lets users visually browse and see their tastes in context; (c) A user happily dancing after Explore More recommended an unfamiliar yet related genre, demonstrating how it aims to broaden perspectives and facilitate personal growth through personalized music discovery.

> *"Now, if you were to do this with other people, you could define the content together and find out, for example, what kinds of things you are missing."* - Student 1

Additionally, the lack of familiarity with translating wellbeing goals into technical requirements or metrics affected designers' ability to thoroughly address wellbeing aims in their solutions. The unfamiliarity with core wellbeing concepts led to doubts about properly executing the methodology. Overall, students lacked confidence evaluating wellbeing considerations throughout the process.

> *"You don't really know when you are doing it right."* - Student 2

Initially, designers faced challenges in evaluating and addressing wellbeing in their design process. However, as the project progressed, the methodology compelled them to consistently test their approaches against the specific context and the people involved. This iterative process gradually sharpened their focus and deepened their understanding. By the project's end, this rigorous application resulted in a notable improvement in their ability to integrate wellbeing considerations, as evidenced by their satisfaction with the

final solutions. This evolution underscores the methodology's effectiveness in facilitating a contextually relevant approach to wellbeing in design.

In terms of method usability, the process involved iterative transitions between research, ideation and prototyping rather than a linear sequence. This is often the case for design processes, however, the framing of the method as steps and the initial visualization of them (not included here) gave them the idea it would be more linear. Further, the overall structure provided helpful guidance, but some inefficiencies occurred due to unclear context definition early on and lack of knowledge about wellbeing. Therefore, significant time was spent exploring literature not directly relevant. To some this may have felt as a waste of time. In reality, it is likely designers will explore directions that in the end of their project may not be relevant anymore—this is what makes them a good designer, being open to multiple avenues and perspectives. The designers suggested to include an introductory wellbeing course, a better explanation the context early on, and to spend more time planning upfront. Despite inefficiencies, designers were satisfied with their process and outcomes overall.

In summary, the Positive AI method demonstrated efficacy in guiding novice designers to translate high-level wellbeing goals into concrete design proposals grounded in user values. Key steps like contextualizing and iterative development focusing on emergent wellbeing priorities were effective, as evidenced by resulting concepts aligned with community experiences. Despite inefficiencies, designers expressed satisfaction that the methodology enabled addressing complex tasks, compelling repeated testing against user perspectives to gradually improve comprehension. This indicates the potential for enhancing efficacy given refinements targeting usability.

In terms of method efficacy key points of attention are:

- Navigating expansive wellbeing literature overwhelmed novices, causing uncertainty grasping concepts to advance confidently. Clearer guidance on scope is required.
- Unfamiliarity translating qualitative aims into technical specifications made operationalization challenging. More extensive scaffolds are needed to aid comprehension.
- Initially lacking familiarity with core wellbeing concepts hampered confidence assessing impacts. But this grew through repeated engagement as understanding increased over time.

Additionally, regarding method usability, certain inefficiencies emerged despite the beneficial structure:

- Unclear initial scope caused detours exploring tangential literature. Signposting priorities earlier would help.
- Absent examples induced difficulty judging step completion. Providing benchmarks would resolve ambiguity.
- Operationalization demands proved taxing for novices. Enhanced support could alleviate strain.
- Shifting between abstract and concrete perspectives around wellbeing felt jarring. Framing this dynamic approach as integral to the design process could smooth transitions.

Having addressed the intricacies of applying the Positive AI method through novice designers' experiences and identified areas for refinement to bolster both its efficacy and usability, we now turn our attention to the last research question. In the following section, we present a narrative-based study involving experts to evaluate the quality of the AI system concepts resulting from the application of the Positive AI method.

5.5. Narrative-based study with experts

The goal of this study is to assess the design quality of AI systems aimed at enhancing human wellbeing. We chose to use a narrative-based study method, following the example of Tromp and Hekkert (2016) who used this approach to analyze a social design method. Narratives can be useful tools for envisioning and assessing the potential impact of emerging technologies that are difficult to prototype or do not yet exist. As Tromp and Hekkert (2016) note, narratives allow people to imagine hypothetical situations as if they were real (Shapiro, Barriga, & Beren, 2010), providing a means to explore near-future scenarios involving novel technologies (Bleecker, 2022). After crafting narratives about not-yet-existent AI technologies, experts can analyze them to evaluate three key dimensions: technical feasibility (could the required algorithms be developed?), business desirability (would companies want to develop this?), and outcome plausibility (could the proposed design plausibly achieve the intended wellbeing benefits?). It is important to consider business incentives when designing AI aimed at promoting wellbeing, since company objectives constrain system behaviors. Without accounting for profit motivations, proposed interventions may conflict with core financial goals. In summary, narrative evaluations allow researchers to imagine and critique the potential societal impacts of emerging AI, while

weighing technical feasibility, business desirability, and the plausibility that intended wellbeing outcomes could actually be achieved.

5.5.1. Method

Procedure and participants

The study involved 17 experts participating in an online questionnaire where they read three narratives describing AI system concepts aimed at enhancing wellbeing. The participants were selected based on their expertise in design, AI, wellbeing, or a combination of these fields. All participants identified as experts in design, 7 as experts in wellbeing, and 10 in AI. They were invited to participate via email.

In the questionnaire, participants first read a narrative envisioning a near-future scenario showcasing one of the AI system concepts. After reading each narrative, they answered two comprehension questions about the concept. Participants then completed a 4-item questionnaire assessing the following dimensions on a 7-point Likert scale ("strongly disagree" – "strongly agree"):

1. "The narrative is realistic and believable."
2. "The suggestion that *[AI system]* promotes wellbeing is realistic."
3. "It would be attractive for a company to develop a platform like *[AI system]*."
4. "It would be feasible for a company to develop a platform like *[AI system]*."

This process was repeated for a total of three narratives. The full questionnaire took approximately 15-20 minutes to complete.

Materials: narrative development

In crafting the narratives, we followed guidelines discussed in Tromp and Hekkert (2016). The narratives were developed by the authors of this paper in collaboration with the students who created the AI system concepts. The ubiquitous contexts of dating apps, food tracking apps, and music streaming platforms provided plausible scenarios while sidestepping charged assumptions. By carefully considering factors influencing perceived realism when designing the final narratives (700-900 words long), the aim was to elicit unbiased evaluations of the AI concepts and their wellbeing claims. The narratives were illustrated with three graphics each that have also been used to visualize the concepts as discussed in section 5.4.2.

A small pilot ($n = 5$) first checked if the three main narratives seemed realistic before the main study, utilizing the Perceived Realism Scale (Green, 2004). No changes were made after this initial pilot. However, minor updates were made after the narratives were copyedited by a native English speaker. A single adapted item ("The narrative is realistic and believable.") was used in the complete questionnaire to assess realism. The narratives can be found in the Appendix.

5.5.2. Results of the narrative-based study

Figure 5.6 presents a graph of the results from the narrative-based study involving expert evaluations across three distinct concepts: MiHue, FoodVibe, and Explore More. The concepts, received moderately high mean ratings on perceived realism, impact on wellbeing, business desirability, and business feasibility. The concepts, on average, received moderately high ratings across the metrics. Specifically, Explore More was rated highest in business desirability, while MiHue and FoodVibe showed similar ratings in perceived realism and wellbeing impact, respectively. Notably, there were no significant differences in ratings among the different expert groups. The associated standard deviations indicate a moderate variation in expert opinions.

On the qualitative front, feedback from two participants indicated that the narratives were lengthy, while another two experts remarked that the stories leaned towards being overtly positive. However, they also noted their understanding of this positive skew, acknowledging the study's context aiming to portray an ideal user experience. Further insights from the expert's feedback will be delved into in the subsequent sections.

5.6. Discussion

This paper introduced the Positive AI design method for developing AI systems that actively promote human wellbeing. Following the framework for evaluating design methods proposed by Cash et al. (2023), we have provided a "chain of evidence" through multiple studies to assess the credibility and robustness of the Positive AI method.

Specifically, we first discussed the motivation for the method based on gaps in current AI design processes. We then explained the nature of the method as a principle-based approach suited for ubiquitous AI systems that seek to actively integrate wellbeing. Next, we detailed the iterative development process applying a cybernetic framework. We then outlined the key steps: 1) contextualization, 2) operationalization, 3) design, 4) implement, 5) reiteration. Finally, we presented evidence for the method's

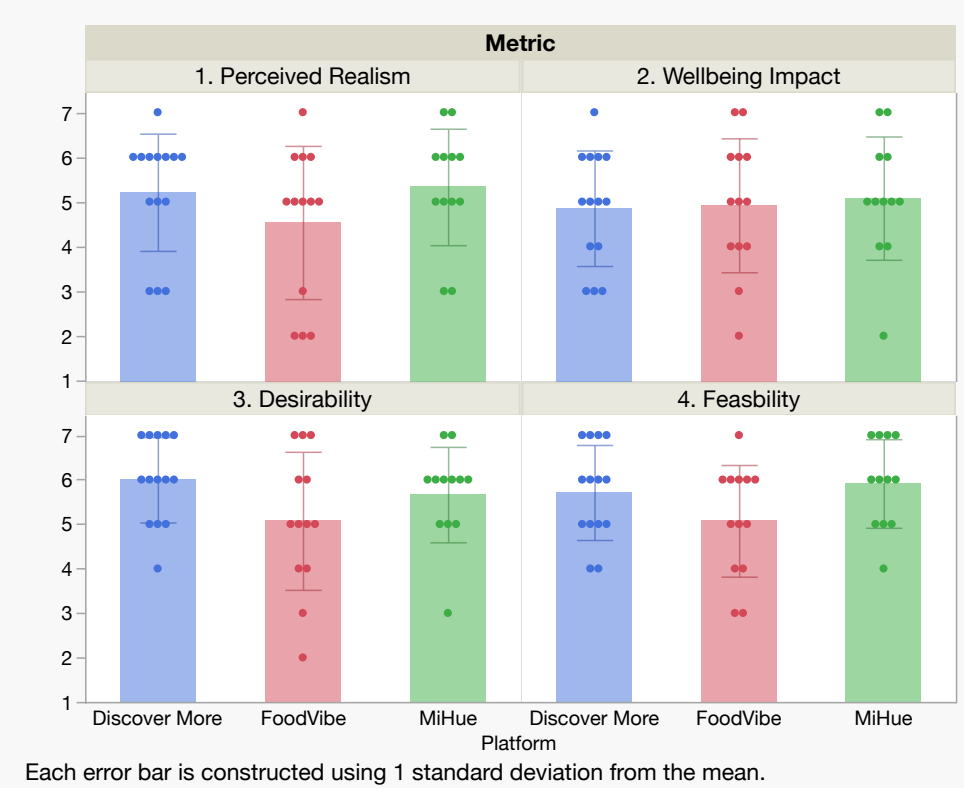


Figure 5.6: Bar chart comparison of expert evaluation ratings across the three concepts: MiHue, FoodVibe, and Explore More. Metrics visualized include perceived realism, wellbeing impact, business desirability, and business feasibility. Each error bar is constructed using 1 standard deviation from the mean.

impact claims through a multiple-case study with novice designers and an expert evaluation assessing the quality of the resulting concepts. Through this initial validation, the method showed promise for improving AI design while also revealing areas needing refinement.

In this final section, we will first briefly reflect on the multiple-case and expert studies, after which we will discuss their limitations. Then, we discuss the position of the Positive AI method with respect to existing frameworks such as VSD and IEEE-7010. Finally, we outline proposed adaptations and future work based on the outcomes of this discussion.

5.6.1. Reflections on the case-study: comparing efficacy and usability

Our exploration through three case studies illuminated how designers integrated wellbeing into AI design, effectively addressing the second research question (RQ2) by uncovering the method's strengths and weaknesses. Specifically, the study revealed a trade-off between the method's efficacy and usability. While the process successfully directed the students' focus towards considerations of wellbeing impacts in their design concepts, demonstrating the method's efficacy, the students also experienced inefficient detours and uncertainty during certain steps of the process. This indicated usability challenges with the current version of the method. For example, the students reported being overwhelmed by the breadth of literature when researching wellbeing theory, causing uncertainty about when sufficient research had been conducted to move forward. Such detours revealed usability issues despite the method's efficacy in eliciting wellbeing considerations. To extend this, Andreasen (2003) emphasizes designers must develop proper mindsets, not just learn procedures, to effectively utilize methods. In this regard, visual models and explanations of theory aid building an effective mindset.

While discussions of usability and efficacy trade-offs are relatively absent in academic literature on design methods, both factors are critical drivers of tool adoption in practice. Farzaneh and Neuner (2019) argue for the indispensability of usability in the effective employment of tools, a point nuanced by Eason (1984), who notes that heightened functionality may inadvertently compromise usability. This insight proposes a balanced view: while deficient usability can significantly impede, it does not outright negate, the potential for adopting efficacious tools. Supporting this, Nielsen (1994) and Blessing and Chakrabarti (2009) highlight that a method's acceptance hinges on both its utility (efficacy) and its ease of use (usability). Consequently, for future studies and method refinements, assessing both these dimensions emerges as a pivotal recommendation.

In conclusion, a deeper comprehension of the interplay between efficacy and usability necessitates not only further method refinement but also the development of guiding heuristics to navigate this balance. The forthcoming section will delve into suggested method modifications designed to optimize this balance, aiming for a seamless blend of efficacy and usability.

5.6.2. Reflections on the expert study: assessing impact, desirability, and feasibility

The expert evaluation study aimed to assess the quality of the AI system concepts resulting from application of the Positive AI method, providing an answer to the third research question (RQ3). Beyond just examining the intended wellbeing impacts, quality was operationalized more broadly through three key dimensions: technical feasibility, business desirability, and outcome plausibility.

While the Positive AI method specifically focuses on supporting wellbeing considerations, results indicated the concepts also scored moderately high on feasibility and desirability. If the method only enabled wellbeing aims, we may have expected substantially lower ratings on the other dimensions. For example, redesigning a platform like *Instagram* for maximum wellbeing benefit may result in features that lock out users for a given amount of time to stimulate physical exercise instead. This would, obviously, not be desirable from a business perspective. Notably, the concepts were not rated perfectly high across all metrics. That is, when inspecting results closely, variation existed both within and across the projects, indicating varying expert perspectives rather than uniform positivity.[13] This variety in feedback underscores the validity of our data and demonstrates that the method can elicit a diversity of responses, highlighting its effectiveness and the nuanced nature of integrating wellbeing into AI design.

Nonetheless, the overall promising ratings for technical feasibility, business desirability, and outcome plausibility imply the method provides useful general scaffolding for creating AI concepts aligned with multiple stakeholder needs. Apparently, by facilitating contextualization and continuous alignment, the process appears to aid in considering diverse perspectives such as users, companies, and engineers. However, it's important to approach these outcomes with caution. The controlled environment of the study may not fully replicate the complexities encountered in real-world applications, necessitating further empirical investigations to validate the method's applicability. Nonetheless, the positive feedback from experts indicates a valuable direction for future research, highlighting the potential of the Positive AI method in advancing the field of AI design.

Similarly, the question arises: would students provided only a general prompt to design for wellbeing, without the structured Positive AI method, have created concepts with much lower quality across feasibility, desirability, and plausibility? A future study comparing outcomes from students given

[13] Data available upon request through the corresponding author.

just a design prompt versus those using the full Positive AI method could provide insight into the degree to which the method specifically enhances these additional quality dimensions beyond wellbeing impact.

In summary, our narrative-based study approach provided valuable insights on both its strengths and limitations, paving the way for further research on its real-world applicability, which we'll explore next.

5.6.3. Limitations of the study

While the Positive AI method offers a promising starting point, extensive future research is required to address current limitations and evolve a comprehensive approach ready for broad adoption. The student projects provided useful initial insights into the application of the Positive AI method by novice designers, as they were able to engage with the method in-depth rather than superficially, helping surface the issues discussed in this article. However, it's important to note that the students received regular coaching and guidance from the supervisory team who created the method, which likely influenced their ability to apply the method successfully and mitigated some of the challenges they faced, such as navigating the breadth of wellbeing literature or translating abstract concepts into concrete metrics.

Moreover, the academic setting differed from professional design environments where designers are presumed to possess extensive domain experience. The temporary nature of the student projects made it difficult to evaluate the long-term, sustained application of the Positive AI approach over multiple iterative cycles. The compressed student timelines allowed only for a single, but thorough, pass through the steps, precluding the examination of how designer understanding and execution might improve through repeated application over many months or years and limiting the analysis of how contextual models develop over time as relationships and priorities shift.

Furthermore, it remains to be seen how the method would be applied by a group of computer scientists who may or may not have a designer on the team and lack access to the authors' expertise. Future work should explore how successful the method is when design teams do not receive expert guidance, possibly by comparing the outcomes of teams given the method with varying levels of training and support. This could provide valuable insights into the method's effectiveness and usability in more realistic scenarios.

To fully assess the long-term, repeated application of the Positive AI method, in-depth research in professional settings tracking multiple iterative cycles over extended time periods is necessary. Real-world validation by professional designers in industry contexts, beyond student projects and fictional cases presented here, is a critical next step. Their feedback would

provide invaluable insights into limitations and areas for improvement when applied in practice. Additionally, some parts of the method can be developed further to provide additional guidance, tools, and examples that make the framework more accessible for practitioners.

Moving forward, real-world validation, additional resources, and studies on longitudinal effectiveness provide key opportunities to address limitations and adapt comprehensive techniques for Positive AI—which we will discuss next.

5.6.4. The method & existing approaches

This research was conducted within the broader context of AI design, a field often challenged by the proliferation of frameworks without sustained dialogue. Therefore, in this section, our goal is to connect our method with existing approaches, thereby advancing the field and identifying opportunities for future work.

The Positive AI Design Method addresses a significant limitation within VSD: the absence of mechanisms for assessing how well values are realized in technology design and the tangible impact of AI systems. This challenge, highlighted by Sadek et al. (2023c), is met by embedding wellbeing assessment as a fundamental principle within our approach. By prioritizing wellbeing, our approach enhances VSD, by making the integration of values within AI systems both measurable and actionable. This focus ensures the operationalization of abstract values into practical design and evaluation criteria, with wellbeing providing a comprehensive and empirical basis for optimizing human values, echoing S. Harris (2010) in the sense that optimizing for wellbeing inherently optimizes for all human values to the extent in which the empirically contribute to human flourishing.

Similarly, the Positive AI Design Method extends and complements the IEEE-7010 standard (Schiff, Ayesh, et al., 2020). While IEEE-7010 lays a robust foundation for the integration of wellbeing metrics into the life cycle of AI systems, our method takes a step further by directly mapping these wellbeing considerations onto existing design approaches. This direct integration ensures that wellbeing is not only assessed as an outcome but actively shapes AI development from the outset. Furthermore, our method extends the IEEE-7010 framework by offering detailed, practical guidance on mapping wellbeing metrics directly to design decisions, thus facilitating a more granular and actionable approach to enhancing wellbeing through AI systems. This approach not only adheres to the holistic perspective advocated by IEEE-7010 but also advances its application by providing a structured method for translating wellbeing principles into concrete design

practices.

Further, this work can be seen as contributing to the broader question of AI alignment, a field primarily concerned with aligning AI technologies with human values (Christian, 2020; Gabriel, 2020). Without going to deep into this area, we recognize an opportunity to advance the field through the methodology presented here. Specifically, the Positive AI method takes a human-centered approach that contrasts with some common perspectives in the field of AI alignment. Much alignment research focuses on technical solutions like reward modeling (Bai et al., 2022; Christiano et al., 2017), and meaningful human control (Cavalcante Siebert et al., 2023). These techniques aim to formally specify values and control objectives that AI systems should optimize for. In contrast, the Positive AI method emphasizes building contextualized understanding of users through participatory research and design processes. It focuses on continuously aligning systems with the multifaceted and emergent nature of human wellbeing through collaboration. In this way, the Positive AI method diverges from alignment approaches that prioritize formal specification of abstract values and control objectives over participatory human-centered design processes. The proposed Positive AI method provides a complementary human-centered perspective to balance the prevalent technical focus in this field by enabling alignment techniques to be *sensitive to human experience.* Still, greater synergy is needed between these approaches to ensure both human values and technical reliability are embedded in mutually reinforcing ways. The Positive AI method's human-centered approach could be enhanced by integrating formal techniques like reward modeling, which may help scale contextual findings.

Lastly, perspectives from explainable AI (XAI) could enhance the Positive AI method. By making transparent how algorithms and data shape user experiences, we can better understand relationships tied to wellbeing (Ehsan, Liao, Muller, Riedl, & Weisz, 2021). In turn, Positive AI's emphasis on establishing causal links between system components and outcomes can progressively demystify the AI system. This increased transparency aids designers in identifying failure points, unintended consequences (Gunning et al., 2019), and can enhance designers' capabilities to co-create with AI (J. Zhu et al., 2018). Furthermore, XAI's focus on addressing diverse stakeholder needs is in harmony with Positive AI, offering techniques to elucidate system behaviors and supporting the creation of AI that nurtures human flourishing (Felzmann, Fosch-Villaronga, Lutz, & Tamò-Larrieux, 2020; Larsson & Heintz, 2020). An exemplary instance of leveraging AI explanations to enhance wellbeing is demonstrated by Hume's EVI, which not only identifies the emotions present in the user's voice but also clearly

indicates the emotions it utilizes for its responses. This approach enriches wellbeing by encouraging deeper, more empathetic communication.

5.6.5. Proposed adaptations & opportunities for future work

Reflecting on our comprehensive evaluation of the Positive AI method, we recognize its significant contributions towards bridging existing gaps in the field, alongside areas that remain open for improvement. In light of these insights, we propose several avenues to move the method forward.

First, providing designers with examples and heuristics may improve method usability. For instance, developing a framework to determine when enough contextual research has been conducted could prevent unnecessary detours. This framework might involve checklists of key relationships or suggested timeboxes. Likewise, heuristics and examples could increase confidence during overwhelming stages such as conceptualizing and operationalizing wellbeing. A recent paper by Peters (2023), synthesizes over 30 years of psychology research to provide 15 of such heuristics that may help technology designers create more wellbeing-supportive user experiences by identifying key areas where AI can significantly impact users' psychological wellbeing, ensuring designs are grounded in well-established principles. Additionally, Fast-paced, preliminary simulations, such as workshops, could acclimate designers to the method and underlying wellbeing theories, offering a practical glimpse into the process and expected outcomes. This would also fulfill a key recommendation from the literature: to promote AI education, ensuring positive impacts and encouraging cross-disciplinary collaboration (Bentvelzen, Woźniak, Herbes, Stefanidi, & Niess, 2022; Morley, Kinsey, et al., 2021; Schiff, Ayesh, et al., 2020).

Further, the current Positive AI method focuses primarily on the design phase of the AI life cycle. This was a conscious choice to better scope the study. Consequently, the method does not yet provide comprehensive guidance for integrating wellbeing principles across all seven stages of the AI development process. Ideally, designers should be involved throughout the entire AI life cycle to ensure the consistent application of wellbeing objectives and to address any emerging challenges or unintended consequences (Norman & Stappers, 2015). Mapping the Positive AI method to the full AI life cycle could improve its real-world feasibility and effectiveness. However, this is a non-trivial task, as it requires further integration and coordination of various perspectives and collaboration among designers and other stakeholders throughout the development process. Future work should investigate how to extend the Positive AI method to the full AI life cycle, addressing the

unique challenges at each stage and developing a comprehensive framework that ensures wellbeing objectives remain at the forefront.

In the same vein, the Positive AI method currently does not comprehensively support the activity of co-designing. While we have identified this as an important gap in the literature and have emphasized co-designing, such as the inclusion of stakeholders, as essential to the success of positive AI, we do not actively discuss how this can be best achieved. Due to the scope and length of this article, this is not the most appropriate place for an in-depth exploration of co-design techniques. Perhaps a future platform detailing the method further could include such resources. For now, we recommend referring to the work of Sanders and Stappers (2008) for general guidance on co-design, and other resources more specifically applied to AI (e.g., Liao & Muller, 2019; Sadek et al., 2023a; Subramonyam, Seifert, & Adar, 2021; Zytko, J. Wisniewski, Guha, PS Baumer, & Lee, 2022). Future research could investigate how these frameworks may complement each other and enhance the Positive AI method.

It is important to notice that the iterative aspect of the Positive AI method relies heavily on frequent engagement with stakeholders, primarily user communities. However, continuous participatory design can be resource-intensive. Investigating efficient techniques for community collaboration at scale, such as those proposed by Peters, Sadek, and Ahmadpour (2023), would strengthen this vital feedback loop and improve the method's real-world feasibility. Developing more streamlined and scalable approaches to co-design will help the Positive AI method better incorporate diverse perspectives and align with the needs and values of the communities it serves.

5.7. Conclusion

This article introduced the Positive AI design method as a concrete approach to develop AI that enhances wellbeing. By centralizing contextual measurement and continuous alignment, it aims to bridge the gap between aspirations and technical specifics. Though initial studies revealed weaknesses, proposed adaptations like examples and heuristics could improve usability while retaining flexibility. Further research should validate sustained applications by experts across iterative cycles. However, by translating wellbeing goals into design practices through principled human-centered techniques, the Positive AI method offers a promising starting point. Moving forward, it provides a foundation to actively shape AI systems that promote human flourishing.

Bonus: Checklist for designing Positive AI

Contextualization phase:

- ☐ Review relevant wellbeing literature and theory
- ☐ Map key components of the AI system (algorithms, interface design, etc.)
- ☐ Conduct qualitative user research (interviews, focus groups, etc.)
- ☐ Synthesize theoretical and user research findings into a contextual wellbeing model

Operationalization phase:

- ☐ Select validated global wellbeing scales
- ☐ Develop context-specific wellbeing metrics linked to system components
- ☐ Ensure metrics enable optimizing algorithms to enhance wellbeing

Design phase:

- ☐ Identify high-potential targets for design interventions via surveys, behavior tracking, etc.
- ☐ Envision system modifications across layers (UX, algorithms) to impact wellbeing
- ☐ Produce artifacts like journeys maps and design principles encoding wellbeing aims

Implementation phase:

- ☐ Guide development process using artifacts from design phase
- ☐ Ensure implemented features align with envisioned optimized interactions

Continuous alignment

- ☐ Regularly re-engage user community via interviews, workshops etc.
- ☐ Revisit contextual model to realign priorities
- ☐ Repeat full process to incrementally enhance wellbeing impacts

6

General discussion & Conclusion

This dissertation was initially motivated by questions around aligning then-prevalent AI systems like social media and streaming platforms to support human wellbeing rather than narrow optimization objectives. At the outset of this research in December 2019, society was in the midst of what some refer to as the "first contact moment" with AI (Harari et al., 2023), represented by widespread interaction with AI through platforms like social media, streaming services, and news feeds. Over the course of my dissertation work, we have entered what could be considered a "second contact moment"—interaction with more advanced AI systems like large language models that are being rapidly integrated across many aspects of our lives. While the AI landscape has evolved significantly since we first conceptualized this research, the core question of how to align complex AI-driven sociotechnical systems with human values and wellbeing remains crucially important. My dissertation focused on identifying effective strategies to guide AI development in a direction that supports human flourishing—conserving a lasting positive effect on society. These strategies aim to contribute to responsible and ethical AI design for the foreseeable future, irrespective of the specific AI techniques used. As we progress towards AGI and superintelligence, ensuring human alignment of AI systems remains an urgent priority if we wish to harness AI's benefits while avoiding potential harm. To uncover strategies and interventions that align complex AI systems with human values and wellbeing, this research was guided by five key questions:

6.1. Research questions

To achieve these objectives, the research has been broken down into the following research questions:

1. How does wellbeing manifest across different AI systems?
   - For instance: *How might a system like Netflix influence various aspects of user wellbeing, such as belongingness or social connections?*
2. How might wellbeing be operationalized in the context of AI?
   - For instance: *What methods could be used to measure the impact of ChatGPT on facets of wellbeing, translating people's lived experience with the platform into measurable metrics?*
3. How might wellbeing-promoting systemic interventions be designed?
   - For instance: *What are the potential areas within the broader ecosystem of TikTok, such as the user interface or curation algorithms, where interventions could be implemented to enhance wellbeing?*
4. How can the impact of interventions on wellbeing be evaluated?
   - For instance: *How might we assess whether a positive intervention in Alexa's voice interaction and response algorithms has the desired outcomes and no unintended consequences?*
5. What are key design steps in designing Positive AI that generalize across contexts?
   - For instance: *Are the processes to operationalize wellbeing in the context of YouTube the same as for Reddit?*

6.2. Summary and contributions to research and design

The research questions above were addressed in this dissertation through several diverse investigations, summarized here:

1. A theoretical analysis proposed a cybernetic framework responding to the need for a systemic perspective when designing AI focused on wellbeing (Section 2.1.1). This contributes to design research and practice by providing a theory describing the mechanism of Positive AI

on wellbeing. The systemic perspective opens up new affordances for design interventions by showing opportunities for influencing wellbeing feedback loops outside of AI algorithms (i.e., within UI/UX elements or other elements in the broader sociotechnical system). It has been published as a conference paper at the *Design Research Society biennale 2022* in Bilbao (van der Maden et al., 2022).

2. This systemic, cybernetic framework informed the identification of seven key challenges for designing positive AI systems and organized the challenges into four main categories: modeling wellbeing, assessing wellbeing, designing for wellbeing, and optimizing for wellbeing (Chapter 3). These challenges can serve as a practical resource for design researchers and practitioners who want to engage with the field of Positive AI. An analysis of these challenges is currently under review for publication at *She Ji: he Journal of Design, Economics, and Innovation.*

3. Based on the idea of cybernetic feedback loops in sociotechnical systems, I designed a large-scale longitudinal assessment of wellbeing (seven iterations, $n = 20{,}311$) to support the needs of over 30,000 staff and students during COVID-19 Chapter 4). Over two years, results from My Wellness Check informed governance at Delft University of Technology by highlighting areas of need and surfacing community ideas for wellbeing interventions. Further, I ran a controlled experiment ($n = 1{,}719$) to compare two validated assessments of wellbeing against my new context-sensitive assessment. This study showed that my assessment was significantly more predictive of overall life satisfaction, which demonstrates the value of contextualizing wellbeing assessments to inform system action, particularly during crisis. These findings were published in *Frontiers in Psychology* (van der Maden et al., 2023).

4. Synthesizing insights from the above investigations, I developed a new design method for Positive AI (Chapter 5). Preliminary versions of the method have been applied within multiple student projects from Bachelor's to Master's level. The method was then applied during three Master's graduation design projects and evaluated in a narrative-based study using expert insights ($n = 17$). Experts rated AI system concepts designed using the Positive AI method as moderately high in perceived realism, wellbeing impact, business desirability, and feasibility. Designers expressed overall satisfaction with the process and outcomes, demonstrating the practical value of the method for

designers. These findings were submitted for publication in *Artificial Intelligence for Engineering Design, Analysis and Manufacturing* and are currently under review.

5. Through an integrated research approach combining theory, challenge analysis, in-context studies, method development, and evaluation, the dissertation uncovers practical strategies for using human-centered design to align AI systems with human values and wellbeing. This work then culminates in a set of recommendations for designers of Positive AI, discussed next.

6.3. Recommendations

Drawing from insights gathered throughout this dissertation, I present a series of recommendations aimed at guiding AI design towards the enhancement of human wellbeing. It's important to clarify that my thesis does not seek to reinvent the wheel. Instead, its novelty comes from bringing together activities of psychologists, designers, policy-makers, and AI practitioners that are typically conducted in isolation from one another. By integrating these disparate strands, I aim to direct them towards a unified objective: Positive AI. This approach underscores the multidisciplinary nature of this work and advocates for a collaborative synergy to bridge the gaps between these essential fields.

The recommendations I put forward are intended for those who aim to design Positive AI and therefore build upon key points from my own work as well as generalizations from the literature. While some of these strategies may already resonate within the broader community, bringing them together here serves two crucial purposes. First, it highlights the importance of a unified effort towards realizing the potential of Positive AI. Second, it outlines a clear and actionable roadmap for future research in this area by bringing together previously siloed contributions. In discussing each recommendation, I will refer to relevant literature and ongoing research efforts. The list below is an overview of the recommendations after which they are discussed in detail in the subsequent sections.

1. **Integrate HCD in AI development cycles**: Designing AI for wellbeing requires a focus on human experience, a systemic perspective, and an iterative approach that integrates qualitative and quantitative assessment. These are key characteristics of the practice of human-centered design.

2. **Balance short-term needs with long-term wellbeing:** Design choices involve tradeoffs between immediate metrics and sustainable wellbeing—a systemic perspective considers this temporal dimension.

3. **Model and measure wellbeing in context:** Define and assess wellbeing in ways that are sensitive to the unique contexts and environments in which AI-human interaction occurs.

4. **Establish multiple feedback loops:** Use qualitative and quantitative feedback loops to bridge gaps between different scales of AI systems—from individual components to broader societal impacts.

5. **Focus on flourishing rather than solely mitigating harm:** The goal of positive AI is to actively enhance human flourishing, not merely minimize risks—this focus goes beyond compliance to guide design choices and processes.

6. **View Positive AI as a continuous journey:** Treat the pursuit of wellbeing through AI as a dynamic process, requiring ongoing assessment, adaptation, and realignment.

6.3.1. Recommendation 1: Integrate HCD in AI development cycles

The first recommendation of this dissertation is the critical importance of integrating Human-Centered Design (HCD) principles in the development of AI systems aimed at enhancing wellbeing. Why is human-centered design so essential to AI for Wellbeing? In short, wellbeing is, at its core, a property of experience. Human-centered design, of all research disciplines, is specifically attuned to human experience and how to shape it. Thus, whether it goes by the name of HCD or otherwise, methods for studying and shaping human experience will be essential for ensuring that AI systems support the nature of wellbeing in experience.

Further, if the goal is to develop AI that harbors human values (Floridi et al., 2018), aligns with the Sustainable Development Goals (Tomašev et al., 2020), and that promotes human flourishing (Ozmen Garibay et al., 2023)—which according to industry leaders it is (Wiblin, n.d.; Zuckerberg, 2018)—HCD methods may be very practical. This practicality has been demonstrated by the activities undertaken in this research—from the participatory design workshops to the iterative development and refinement of My Wellness Check. These empirical efforts point to pathways by which HCD facilitates a deeper understanding of the experiential dimensions of

wellbeing, enabling the design of AI systems that are not only technologically sophisticated but also deeply resonant with the nuanced contours of human life.

As such, synthesizing insights from both the literature and dissertation research, HCD methods can contribute in the following ways:

1. **Attention to human experience:** This quality emphasizes the importance of understanding and valuing human experiences in AI design, recognizing that wellbeing is fundamentally experiential. It involves a commitment to aligning AI technologies with the real-world emotional and behavioral impacts they have on users.

2. **Systemic perspective**: HCD adopts a holistic view, considering the broader system in which AI operates. This includes understanding the interactions between different components and stakeholders in the AI ecosystem. Embracing this systemic perspective reveals the significance of including diverse views and interests perspectives and the ability of each component in the AI ecosystem to drive positive change.

3. **Iterative improvement**: HCD involves an iterative design and development process. This means continually testing, refining, and improving AI solutions based on feedback and changing requirements

Attention to human experience. Research methods based on assessing human experiences are essential to any wellbeing optimization approach, whether in AI or otherwise. This logically follows if one recognizes that wellbeing is, first and foremost, experiential in nature. To take an example extensively discussed in this dissertation, Facebook sought to promote wellbeing by changing their algorithms to boost "meaningful social interactions" (MSI) (Stray, 2020). Despite being well-intentioned, these changes instead accelerated misinformation and the experience of outrage as publishers exploited the algorithmic changes to optimize divisive clickbait (Hagey & Horwitz, 2021). This suggests that Facebook's metrics were misaligned with their wellbeing objectives. A more effective approach would have been to align these metrics with comprehensive assessments of human experience, instead of narrowly as increased interactions between friends and family (Mosseri, 2018). By adopting a more comprehensive approach to assessing human experience, Facebook could have gained valuable insights into which wellbeing factors are affected *together with* wellbeing fluctuations. Identifying these factors would help pinpoint potential areas for intervention. With these areas in mind, they could then set up targeted experiments

to determine which specific parts of the system *cause* these fluctuations. Incorporating HCD principles into this process could help ensure that the definition of MSI encompasses users' authentic experiences and wellbeing, rather than just algorithmic interactions.

While ongoing work indirectly addresses human experience by optimizing for values such as fairness and responsibility, efforts that directly focus on capturing and enhancing the nuanced aspects of human experience remain limited. Nonetheless, promising examples exist such as the initiative from Anthropic and Collective Intelligence Project (2023) who align the values of their AI systems to public opinion, Spektor et al. (2023) who consulted workers to improve algorithmic management for wellbeing, and Stray and Hadfield (n.d.) who were able to conduct a controlled trial on Facebook's platform to prioritize long-term value for users instead of immediate engagement metrics.

Systemic Perspective. Embracing a systemic perspective yields two important realizations. First, it necessitates incorporating a range of diverse perspectives. This is crucial because AI's implications span multiple disciplines, and as demonstrated in previous examples, it's essential to involve all relevant stakeholders, especially those most affected, for effective AI design. Second, a systemic perspective enables the opportunity for interventions to take place across all layers of the system, from the algorithmic level and UI/UX design and even to broader ecosystem interactions outside the platform. Thus, rather than relying on corporations to implement changes in their platform or algorithm, the perspective presented here is that a human-centered design approach can support AI for wellbeing through many other design interventions.

Some more examples may be helpful at this point to help show how a systemic view can support HCD efforts.

1. Based on HCD research, a small design firm could create an add-on for Facebook that promotes more meaningful real-world interactions. This add-on could facilitate organizing community events or social projects, encouraging users to engage in meaningful activities beyond the platform. Or, perhaps there is no technological development at all—it might just be a technique for using existing tools to have better & more aligned outputs.

2. Based on HCD research, people may hypothesize that Netflix could promote better social engagement. Then, to act on the research, designers might produce a website to promote the formation of "Netflix

watch clubs" to encourage people to expand their horizons and enable meaningful conversations with friends, family, and colleagues based on their viewing experiences. Such an approach fosters social connections and alleviates the guilt associated with binge-watching. Critically, it is an AI for wellbeing design activity that does not rely on Netflix itself to make changes to its algorithm or platform.

3. As another innovative use of AI for wellbeing, I can point to our own efforts in organizing "Cocktails & AI" community events where participants at TU Delft gained early experience with tools like the *OpenAI Playground* and *Stable Diffusion* while enjoying a nice cocktail. This setting fostered positive relationships and a sense of belonging among attendees, simultaneously enhancing AI literacy. This exemplifies that when AI impacts communities, every member can feel empowered to initiate action and even to design new systems.

To clarify, while AI for Wellbeing can be initiated by individual designers or communities through the design of interventions, this does not absolve companies of their responsibilities. Nevertheless, designers need not wait for corporate actions to make a difference and can take steps to effect change on their own through third-party interventions as those mentioned above.

Suggestions for adopting a systemic perspective in AI design, while not entirely new, are still emerging and evolving within the field (Dobbe et al., 2021; Sadek et al., 2023b; Sartori & Theodorou, 2022; van de Poel, 2020; Vassilakopoulou, 2020). The current dissertation contributes to this growing body of work by providing concrete examples of how such a perspective can be applied in practice, particularly highlighting the myriad of interaction opportunities that a systemic view unveils (see Table 4.4 and Section 5.4 for detailed discussions).

Iterative Improvement. In designing complex systems that involve both technology and society, it is often seen as more advantageous to generate numerous small ideas rather than implement sweeping changes abruptly (Norman & Stappers, 2015). This approach enables a meticulous evaluation of how each minor adjustment influences overall wellbeing—concentrating on the incremental, considerate improvements that circumvent major disruptions, in order to guide users towards a primary objective of enhancing wellbeing.

Spotify's 2023 strategy exemplifies this principle. They methodically introduced significant changes but in a subtle, layered fashion. Their updates focused on bolstering user engagement and community interaction—but were released incrementally to allow for a gradual adaptation (Spotify,

2023). Features like enhanced artist-fan connection tools (e.g., recorded video messages) and individualized music discovery options were rolled out over multiple stages. The incremental nature of the change permitted Spotify to gauge user reactions and make necessary refinements, ensuring a seamless integration that resonates with user preferences and cultivates a vibrant, interconnected community experience. While this recommendation may be targeted more towards those more capable of making high-level interventions, it should aid in the realization that positive change does not come from wholesale redesigns *per se* but can come by making small incremental changes as demonstrated in Chapter 4.

HCD's focus on iterative prototyping and integrating diverse perspectives is vital for adapting AI to emerging wellbeing needs. The two upcoming sections will address the previously mentioned aspects of context-specific modeling and detailed recommendations for implementing these approaches by establishing multi-layered feedback loops.

6.3.2. Recommendation 2: Balance short-term needs with long-term wellbeing

The next recommendation is to be mindful of balancing immediate with future needs. As discussed in Chapter 3, and demonstrated in Chapter 4, there is a dichotomy between AI's proficiency in meeting immediate desires and its challenges in addressing long-term wellbeing. Algorithmic optimization cycles can quickly adapt to "what you want" in the moment, rather than "what you need" in the long term.[10] This mismatch of *pace* (short-term and long-term) is a fundamental challenge for designers of AI for wellbeing (as discussed in Chapter 3). To navigate this dichotomy effectively, the following sub-recommendations are proposed:

1. **Understanding the pace mismatch:** AI systems must adapt to the dichotomy between people's immediate desires and long-term wellbeing needs by developing mechanisms to assess the interplay of short and long-term impacts on flourishing.

2. **Putting in place mechanics to assess:** Employing context-sensitive metrics alongside recognized wellbeing measures enables a comprehensive understanding of individuals' experiences to determine optimal scenarios balancing immediate and long-term interests.

[10] It should be clear that "need" here refers to fundamental need satisfaction (Desmet & Fokkinga, 2020) rather than needs in the sense of simple "wants" (Papanek & Fuller, 1972).

3. **Managing tradeoffs between short-term and long-term:** Managing tradeoffs between people's short-term desires and long-term wellbeing is essential for Positive AI; rather than an admonishing "AI nanny," future research around aligning immediate and fundamental interests could enable *wiser* applications.

Understanding the pace mismatch. For AI to effectively promote wellbeing, it must consider the immediate impact and how short-term interactions interact with long-term wellbeing. Namely, indulging impulsive wants countermands wellbeing because immediate gratification can lead to negative consequences in the long term (Desmet & Fokkinga, 2020). When individuals give in to their impulsive desires, this can undermine their overall wellbeing if it leads to the neglect of more important (but less immediately desirable) goals and responsibilities. The key challenge lies in understanding this dynamic—how today's actions influence tomorrow's wellbeing. Engaging directly with users to uncover their life priorities and areas where AI can have a tangible impact is crucial. This introduces a dilemma related to the earlier discussion on HCD: Are HCD methods that focus on understanding "today's" experiences enough to predict "tomorrow's wellbeing?" While I do not possess a *Magic 8 Ball*, it is possible to implement methodologies that adjust to changes in wellbeing over time, a topic we will explore in upcoming recommendations. Acknowledging the need for understanding the relation between a system's capabilities for acting in the short-term and their long-term wellbeing effects, the next step is to put in place mechanisms that can causally couple system actions to delayed effects.

Putting in place mechanics for coupling effects. Understanding the relationship between short-term impacts and long-term wellbeing necessitates the implementation of specific mechanisms that can establish a causal connection between these two dimensions. This objective can be achieved by integrating local indicators with global wellbeing measures. For example, a general question such as "On a scale of 1 to 10, how satisfied are you with your life?" becomes more meaningful when complemented with localized, context-rich insights—as extensively demonstrated in Chapter 4. This twofold approach gives due recognition to immediate personal experiences while ensuring alignment with the broader spectrum of long-term human flourishing. The implications of this approach, which involves modeling wellbeing in context, will be discussed further in Recommendation 6.3.3. With these mechanics in place, we may be able to develop an understanding of what the "best" scenario is in a given situation. Yet, the question remains whether people find these best scenarios desirable.

Managing tradeoffs. People often want things they know are not good for them, such as smoking cigarettes. Jenny Holzer's phrase, "Protect me from what I want," succinctly captures this tension. However, the thought of an admonishing "AI nanny," constantly telling us what is good or bad for us, is not necessarily an appealing vision of Positive AI. There may be room for balance—e.g., "wellbeing cheat days"—to enjoy short-term pleasures guilt-free. Yet Holzer's phrase keeps resonating, encapsulating the dilemma between what we want and what supports human flourishing. This tensions and the fact that it is not straight-forward to design interventions to enhance wellbeing has been extensively demonstrated throughout this dissertation. As such, further research around aligning immediate and fundamental interests could enable *wiser* applications.

Again, talking to people and understanding their broader experience over time remains fundamentally important. But this makes clear that "talking to people " does not simply mean asking them what they want and running with it—as has long been known in the context of HCD (Norman, 2005). Instead, it implies engaging in a more in-depth and empathetic conversation to uncover the underlying motivations, emotions, and aspirations that drive people's choices and preferences.

In this vein, Tromp and Hekkert (2019) offer insightful strategies for managing tradeoffs between conflicting values in the context of social design: 1) Transform: integrating societal issues into personal experiences, thereby making broader concerns directly impactful to individuals; 2) Bypass: changing the perception or experience of behavior to connect with alternative interests, leading to indirect societal benefits; and the most desirable among these 3) Resolve: creating design solutions that fundamentally shift the behavioral context, resulting in behaviors that naturally align with societal wellbeing. For the latter, they provide the example of a traffic light that counts down until the light turns green. This aligns short-term objectives (managing a person's impatience) with long-term objectives (traffic law obedience). This extends to the context of Positive AI in the sense that ideally, the AI system can be designed in a way that naturally supports contextual wellbeing—aligning short- and long-term objectives. For example, an AI assistant that schedules meetings could resolve the dilemma between individual preferences for timing and the collective concern for work-life balance. By factoring in historical data on optimal energy levels at certain times of day, it would automatically suggest meeting times aligned with both productivity and healthy work boundaries.

Implications for AI alignment. This discussion also raises critical questions about definitions of AI alignment. OpenAI's approach aims at

aligning AI with human *intent* (OpenAI, 2022). However, our intentions do not always align with what is best for us. Returning to the car repair scenario from the introduction (see 2.1.3), what if the owner specifically asks the mechanic for minor servicing like "change my tires and oil," but after inspecting the car, the mechanic suggests additional necessary repairs for the car's longevity, racking up further costs? A conundrum occurs between the owner's desire for a low bill and the car's future need for comprehensive repairs. There is a misalignment between the owner's specific request and the mechanic acting in the owner's best interest. Depending on the owner's financial context, this tradeoff requires resolution—can they afford the ideal repairs, or will extra costs risk debt (future misalignment)? This highlights how agent-principal misalignment can arise even from benevolent intentions for another's ultimate good. In an AI context, one can imagine wanting simple entertainment from technology (i.e., a little binge-watching on Netflix) while this temporary comfort conflicts with long-term thriving.

6.3.3. Recommendation 3: Model and measure wellbeing in context

To design Positive AI, wellbeing must be defined and operationalized in context. This means complementing universal, globally validated theories and measurement scales with locally situated understandings and indicators of wellbeing. Chapters 4 and 5 demonstrate how established global models of wellbeing such as PERMA (Seligman, 2010) or PWB (Ryff, 1989) can be coupled with local indicators and adapted through participatory processes that elucidate conceptions of flourishing unique to impacted contexts. This process leads to a ***contextual model of wellbeing*** which can be operationalized and eventually designed for.

Chapter 4 extensively discusses how, during the My Wellness Check project, it became clear that existing theories of wellbeing and their respective assessment instruments were insufficient in providing stakeholders with the necessary information to take responsive action and support wellbeing during the COVID-19 pandemic. As the context of COVID-19 presented completely new challenges to wellbeing, these were not yet adequately represented in the literature. Consequently, we had to contextualize existing frameworks and operationalize them to identify potential areas for positive intervention. This real-world example directly aligns with the discussions of others (e.g., Kross et al., 2021; Stray, 2020), who have highlighted that emergent contexts and technologies require the adaptation of existing wellbeing instruments and the construction of contextual understandings of wellbeing. For instance, Kross et al. (2021) suggest that the lack of

consensus regarding social media's impact on wellbeing may be attributed to the insensitivity of existing instruments to these novel contexts, such as social media platforms.

Thus, modeling and measuring wellbeing in context provides designers of Positive AI with the following five critical factors for success:

1. **Actionable wellbeing data.** Actionable data that combines specific wellbeing metrics with qualitative insights is essential for designing effective interventions that support user wellbeing.

2. **Attributing wellbeing fluctuations.** Attributing wellbeing changes to specific interventions requires contextualizing wellbeing by linking local indicators and global metrics for dynamic, situationally-appropriate support.

3. **Adaptability to contextual shifts over time.** AI for wellbeing requires adaptable, context-sensitive measures that capture evolving notions of human flourishing, aligning metrics and interventions with cultural transformations and emerging technological impacts.

4. **Awareness of Goodhart's Law.** To avoid that metrics drift, designers should not treat measurements as objectives in themselves but as fallible proxies for higher aims like human flourishing.

5. **Use theory to inspire ideas in context.** Translating between quantitative system metrics and qualitative insights from people's lives is essential to contextualize wellbeing when translating between literature and lived experience.

Actionable data. In the history of psychology, the purpose of operationalizing and assessing wellbeing is to better support the psychological study of the phenomenon of wellbeing (Diener, 2019). Examples exist of psychologists such as Lyubomirsky and Layous (2013) using their research to directly inspire intervention, yet most theories were not originally developed for this purpose. As such, even excellent measures of wellbeing do not necessarily help to inform specific design interventions. For instance, the *Life Satisfaction Scale* is one the best-validated measures of wellbeing (Cooke et al., 2016; Linton et al., 2016)—yet, this measure has little to say that can help turn collected wellbeing assessment data into actions that support wellbeing improvement. 'Life Satisfaction' is so broad that any single design intervention is unlikely to "move the needle" on a measure of someone's

wellbeing nor does a life satisfaction score alone point to specific areas for intervention.

Assessing more specific factors of wellbeing results in more actionable data in the sense that it gives designers information about where in the system they may be able to intervene. For instance, in the My Wellness Check project (Ch. 4), we employed a slate of global wellbeing measures related to life satisfaction, mood, physical health, and motivation, and combined them with local indicators related to their student-life experience, home-working space, substance intake, and academic experience. This approach was further enhanced by integrating detailed open-ended questions. Such a combination provided us with specific insights for intervention areas (see Table 4.4). This method helped in identifying concrete concerns or opportunities, allowing for a more targeted and meaningful analysis of the data.

Consider a scenario where analysis reveals a correlation between increased depression indicators among teenagers and more time spent on social media (Keleş et al., 2019). While this finding connects depression to social media usage, it lacks actionable insights. Coupling this quantitative data with qualitative research methods may, for example, uncover unmet needs related to community belonging, self-expression, body image, manifesting as emotional distress and depression (Hagey & Horwitz, 2021). These insights can guide designers in developing targeted interventions, such as features that facilitate healthy peer connections, to alleviate the observed increase in depression.

This example demonstrates the importance of using contextualized metrics that link numerical data to specific system opportunities, allowing positive AI developers to act on the signals provided by wellbeing assessments. Quantitative measurements alone often fail to guide tangible improvements without the context of users' priorities and experiences. Incorporating localized narratives and qualitative insights that reveal what matters to users makes the data actionable, enabling designers to create effective, targeted interventions that support user wellbeing.

Attributing wellbeing fluctuations. Wellbeing is a complex, multi-faceted phenomenon highly susceptible to external events. Attributing changes in wellbeing to a design intervention or specific system component requires establishing causal links between them (Fokkinga et al., 2020). Context-sensitive measures are instrumental in discerning which factors fluctuate in tandem with wellbeing. By identifying this relation, designers can run controlled experiments aimed at causally connecting these fluctuations to particular aspects of the system they are designing for. By meticulously tracking how different elements impact wellbeing, designers can refine their strategies to ensure interventions are effective and affirm that Positive AI design choices directly contribute to human flourishing.

Say that Facebook introduces a new wellbeing-supportive feature, such as prompting users to take breaks after a certain amount of time spent on the platform. To determine whether this intervention is effective, Facebook could couple global wellbeing metrics (e.g., life satisfaction) with context-sensitive measures directly related to the feature. For instance, they might track the frequency of users taking breaks when prompted, monitor changes in user-reported emotional states before and after breaks, and assess fluctuations in time spent on the platform. By analyzing these localized indicators alongside broader wellbeing measures, Facebook could establish a causal link between the new feature and improvements in user wellbeing, as opposed to attributing the changes to external factors like shifts in personal circumstances. Similarly, in the My Wellness Check project, distinguishing the impact of our interventions from external influences on wellbeing, such as being able to go outside again due to lifted lockdowns, required a combination of global and context-specific measures.

In short, this process of contextualizing wellbeing by coupling local indicators with global metrics enables mapping relevant system components to specific wellbeing facets. This approach allows for dynamic, situationally-appropriate support and helps *attribute* wellbeing fluctuations to Positive AI interventions.

Adapt to contextual changes. As contexts change over time, AI for wellbeing needs context-sensitive measures of wellbeing that can also adapt to the changing circumstances. As external events happen and societal priorities for wellbeing shift, contextualized assessments can capture evolving notions of human flourishing (and languishing). For instance, the COVID-19 pandemic radically reshaped conceptions of wellbeing almost overnight. Early priorities around physical health and safety gave way to pressing mental health, economic, and social connectivity concerns as the crisis

endured. Systems leveraging contextualized feedback could pivot features and recommendations to align with these shifting realities. This involves continually recalibrating operationalizations based on on-the-ground truths about people's lived experiences. Doing so allows metrics and interventions to align with cultural transformations and emerging technological impacts on localized wellbeing. In essence, contextualization enables flexibility to support communities' dynamic conceptions of wellbeing as they progress through major events that reshape perspectives, gain new insights that deepen understanding, and develop emerging collective values that transform social norms.

Awareness of Goodhart's Law. Goodhart's law states that "any observed statistical regularity will tend to collapse once pressure is placed upon it for control purposes" (Goodhart, 1975, p. 96). Essentially, when metrics are treated as objectives in themselves (versus as measures of the objective), they can become distorted and fail to accurately represent the underlying phenomenon of interest. This is a common issue in AI (Thomas & Uminsky, 2020) and highly relevant for Positive AI because measurement models guiding the design process can be prone to gaming or unintended consequences if given too much weight.

To avoid potential adverse effects from Goodhart's law, designers should view wellbeing metrics as contestable and not the final word. Measurements are helpful signals to provide insight into a system, but they are limited in how well they can fully capture complex human experiences. The ultimate objectives should remain grounded in the higher aims for the target (e.g., human flourishing) rather than simply improving a metric. Maintaining the distinction between a visionary objective and the specific measurable goal may lessen the chance of gaming or distorting metrics. For instance, if an LLM-based wellbeing coach measures its success solely by the number of interactions it has with users, people might engage with it more frequently to fulfill this metric. However, this increased engagement doesn't necessarily indicate improved mental wellbeing or effective support. Users might interact more just to boost the metric rather than gain real value from the conversations, leading to a misrepresentation of the AI's impact on their wellbeing. Instead, the designers could have supplemented the usage metric with qualitative user feedback and outcome-based wellbeing assessments to ensure the interactions were actually beneficial.

Use theory to inspire ideas in context. To contextualize wellbeing, translating between literature and lived experience is key. The intent is not to devalue existing wellbeing research but rather to critique and build upon

it in conversation with people impacted by the daily systems they interact with. These stakeholders can illuminate which facets of wellbeing manifest in their context, guiding exploration. Most importantly, they identify under-examined areas needing elucidation. For example, a practical approach to enrich this exploration is to use model factors as brainstorming prompts. By asking, "How might we improve *[a specific wellbeing factor]* in our current designs?," we can leverage theoretical models as a foundation for creative thinking while ensuring our interventions are deeply rooted in the actual experiences and context of those affected. In essence, feedback channels must translate between established theory and on-the-ground experience. This enables oscillation between qualitative insights from people's lives and quantitative system metrics to contextualize wellbeing. Establishing such feedback loops is essential for positive AI design and can take many forms, as discussed in the next recommendation on bridging system scales.

6.3.4. Recommendation 4: Establish multiple feedback loops

To design AI for wellbeing, one must connect multiple feedback loops across several parts of the system. Specifically, a connection has to be made between different loops of assessment (qualitative and quantitative assessment, self-report, and behavioral metrics), between stakeholders (designers, management, the user community), and across pace layers (i.e., short-term desires and long-term effects, as extensively discussed in recommendation 2).

Assessment Feedback Loop: Are the quantitative measures and metrics aligned with qualitative experience?

Qualitative insights are essential for informing quantitative system metrics and for supporting continuous alignment. System metrics (derived from clicks, likes, choices, etc) are highly optimizable in an algorithmic manner. However, these are only meaningful insofar as the metric being optimized connects to the qualitative experience of constituents of the system. For instance, increased time on task might be a metric for engagement; however, it could also be representative of frustration (i.e., when tasks are confusing). Only qualitative insights from human experience can serve as the "ground truth" for the meaning of these metrics—and the connection between the metrics and the experiences need to be aligned over time (as they can drift, see van der Maden et al., 2022).

Transitioning between lived experiences and system metrics is not a straightforward exercise. Human-centered methods can be used to understand user experiences, which can then inform local indicators that can be

used as metrics for system success. A concrete example of this process is discussed in Chapter 4, where this process occurred with each iteration. For example, students' experiences of social isolation during remote learning were translated into survey questions about their sense of belonging and community. The percentage reporting feeling socially disconnected became a metric that informed organizing events to improve social wellbeing. Likewise, students' challenges with homework environments were quantified through survey measures of their dissatisfaction with factors like ergonomics—resulting metrics showing areas of need motivated by providing ergonomic equipment to improve these experiences. Through iterative community-led design and analysis, lived experiences were systematically translated into actionable metrics. This created an optimization loop between assessing wellbeing needs and targeting them through institutional and grassroots initiatives. Maintaining tight coupling via community-led design builds trust and motivation for administrators and other university stakeholders to act upon this data.

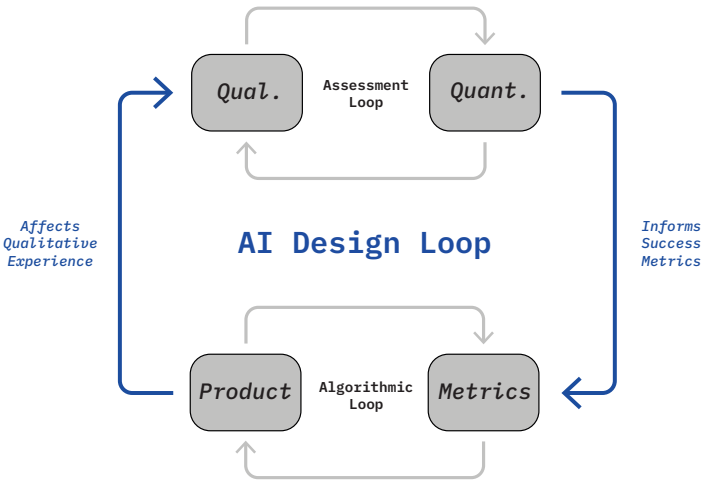


Figure 6.1: AI Design Loop – Conceptual flowchart depicting the interplay between qualitative (Qual.) and quantitative (Quant.) feedback, and how it can inform an algorithmic loop. The directional arrow indicates the process by which qualitative insights inform the development or refinement of quantitative measures, illustrating the cyclical feedback mechanism through which HCD can contribute to better success metrics for AI.

In short, connecting qualitative experience to quantitative measurements enables the development of instruments for research at scale (e.g., surveys), which then informs the design of system metrics and facilitates the subsequent testing and refinement process (optimization), see Figure 6.1. You might, for example, conduct qualitative research on Instagram and discover that body image is a significant concern among teenagers on the platform. This formative research could highlight issues around social comparison—how often teenagers compare their bodies or lifestyles to peers they see on Instagram. Consequently, developing a metric to quantify social comparison rates among this demographic can provide useful contextualization. For example, measuring how many edited versus non-edited images a specific demographic (i.e., teenagers) sees per browsing session. Explicitly defining metrics based on salient qualitative insights allows for thoughtful quantification of complex user experiences.

Algorithm Feedback Loop: Is the product supporting the improvement of metrics?

On the one hand, an AI system itself undergoes constant development, driven by metrics that algorithmically update the product. This process ensures the system adapts and improves in real-time, staying aligned with user preferences and behaviors. On the other hand, these metrics serve a dual purpose; they also inform human designers about key aspects of user engagement and satisfaction. This dual feedback mechanism, where metrics guide both algorithmic refinements and human-led design changes, creates a synergistic environment. It raises a pertinent question: Is the product effectively supporting the continuous improvement of these metrics? This feedback loop is essential for fine-tuning the product to meet user needs better, leading to a more efficient and user-centric recommendation system.

AI Benchmark Feedback Loop: Is the AI meeting benchmarks for success?

Academics may not have much control over how AI systems are developed, but they can influence how they are evaluated. That is, academics can create and apply benchmarks, which serve as standardized methods for testing and measuring the performance of AI systems. These benchmarks are essential tools for evaluating and comparing different AI models, ensuring a consistent and objective assessment of their capabilities. This involvement establishes a feedback loop between the academic world and AI developers, empowering academics in their ability to steer AI innovation. The importance of this role is highlighted by the recent successes of systems like *Gemini* (Google) and

Apache 2.0 (Mistral), which gained significant attention for outperforming earlier models in their class on benchmarks like MMLU (Hendrycks et al., 2020) and GLUE (A. Wang et al., 2018). That is, benchmarks not only foster improvements in AI but also offer a means to gain publicity, which is crucial for securing funding and furthering scientific research. Thus, while academics might not directly control AI development, their influence in the evaluation process plays a critical role in shaping the progress and perception of AI technologies.

Further, as systems evolve (also referred to colloquially as GPT-n+1), these benchmarks become increasingly important. There's a growing demand for benchmarks that are comprehensive and address broader aspects of AI performance. We will need more contextual, complex evaluations that probe ethical reasoning, biases, value judgments, and the capability to explain decisions sensibly (Burnell et al., 2023). New benchmarks and query types will be imperative, like a toxicity benchmark that measures an AI system's tendency to make harmful, biased, or unethical statements (Lynch, 2023). While traditionally not a primary focus for human-centered designers, HCD can significantly contribute by ensuring that benchmarks are technically sound and aligned with human values and experiences. One could imagine a process where community stakeholders actively participate in defining the benchmarks, contributing their diverse perspectives and expertise. This collaborative effort could involve end-users, ethicists, sociologists, and other relevant parties in a dialogue to identify key aspects of human experience that should be captured in the "community-driven benchmarks."

This introduces an intriguing question: What types of benchmarks align best with assessing wellbeing, and how can we effectively benchmark outcomes such as wellbeing that aren't immediately measurable? Future research should focus on investing resources in this area, aiming to devise new methods and benchmarks that can accurately evaluate the influence of AI systems on human wellbeing.

Business Feedback Loop: Is optimization supporting business goals?

For many companies developing AI systems, quarterly financials, and shareholder returns likely to remain the foremost priority. Focusing on short-term business growth can conflict with pursuing long-term wellbeing objectives, as decisions optimizing immediate revenue may undermine wellbeing over time. To address this, some emphasize managerial optimization through incorporating wellbeing metrics into business reviews (Morley, Elhalal, et al., 2021; Schiff, Ayesh, et al., 2020; Stray, 2020). However, typical quarterly reporting cycles mismatch the longer-term timeline needed to demonstrate

wellbeing impacts. Overcoming this may require researchers to produce interim deliverables with business-relevant indicators to communicate the value of wellbeing interventions effectively. Nonetheless, this tension between financial incentives and social impacts is less relevant for start-ups or grassroots initiatives happening external to corporate platforms. External researchers and designers have greater flexibility to develop interventions promoting wellbeing in the absence of pressures for immediate profit.

User Experience Feedback Loop: Is the product aligned with user experience goals?

This is a common type of loop encountered in AI and various corporate settings. Yet it remains relevant to be cognizant of this as wellbeing objectives shouldn't compromise UX objectives. Consider, for example, the earlier mentioned "AI nanny," if an intervention is implemented that is perceived as undesirable yet good for wellbeing, people are not likely to use it or even may stop using a platform because of it. It can even be argued that undesirable user experiences directly go against wellbeing, at least at a hedonic level.

For contemporary AI companies, common tactics include creating digital forums for open-ended user input (e.g., platforms like *Discord* or *Slack*); conducting surveys to gather structured data on user needs; convening focus groups or participatory workshops to enable co-creation and sensemaking; and observational studies to analyze emergent community behaviors and pain points. The ideal approach combines qualitative and quantitative data gathered through multidimensional channels to enable bi-directional learning between system designs and users. For instance, in Chapter 4, we have seen the efficacy of engaging stakeholders of all system layers in the analysis of large amounts of qualitative data. Through workshops, these were translated into actionable insights and usable metrics, but they also instigated action at a local level as the collaborative analysis events caused participants to be reflexive and apply insights in their own context.

In summary, various creative bridging techniques have arisen that can be tailored and combined to fit specific organizational contexts and system maturity levels. Indeed, significant challenges remain in achieving meaningful cross-scale translation for massive platforms like Netflix, which interface with hundreds of millions of users. However, promising examples are emerging from "smaller" scale (i.e., 15 million users) sociotechnical systems like the *MidJourney* community. This system actively gathers user input via *Discord* to participate in community discussions and surveys to improve system outcomes related to creative expression iteratively.

6.3.5. Recommendation 5: Focus on flourishing not merely mitigating harm

Mainstream AI design conventionally emphasizes harm reduction—ensuring systems do not adversely impact users. This is also prevalent in ethical AI literature, which emphasizes non-maleficence (Floridi et al., 2018) and the avoidance of harm (Ozmen Garibay et al., 2023, Table 1), as discussed in Chapter 1 & 3. However, avoiding harm does not guarantee wellbeing. Positive AI builds on Positive Design's strength-based approach of enhancing lives rather than solely removing negatives: "The process of designing for [wellbeing] is different from a traditional problem-focused design process. Therefore, the design field needs approaches that fit with this new vision and the intention to focus on opportunities enabling people to thrive and creating a lasting effect on people's lives." (Desmet & Pohlmeyer, 2013, p. 2)

In other words, Positive AI warrants shifting focus from focusing on problems to proactively cultivating opportunities for human wellbeing. This diverges from the typical mindset in AI development that prioritizes technological achievements. Positive AI proposes more proactive, iterative design that is sensitive to potential downstream effects rather than reactive approaches that passively await consequences.

Actively promoting wellbeing. For example, an intervention aimed at promoting wellbeing could involve a feature that encourages educational and personally enriching content. This could be a personalized recommendation system that tracks the user's time spent on the platform and analyzes their viewing habits to suggest educational and uplifting content tailored to their interests. For instance, if a user frequently watches entertainment videos, the system could intersperse these with educational content, perhaps related to their watched topics, to create a more balanced and enriching viewing experience. This approach goes beyond merely avoiding harm; it actively seeks to enhance the user's knowledge, curiosity, and personal growth, aligning with the principles of Positive AI.

Unintended consequences. However, adopting this Positive AI approach, while advantageous, brings its own set of nuanced challenges. For instance, while the system promotes educational content, creating an echo chamber effect is risky. If the AI continuously reinforces a user's existing beliefs by only recommending similar content, it may inadvertently limit their exposure to diverse viewpoints and hinder critical thinking. Additionally, the algorithm's decisions on what constitutes 'educational' content could be influenced by underlying biases, potentially leading to a skewed representation of

information. This emphasizes the need for nuanced and creative design solutions, iteratively implemented and reevaluated.

Harm-mitigation for wellbeing. Harm mitigation and Positive AI are complementary, with harm reduction also contributing to increased wellbeing. Take, for example, a hypothetical feature on YouTube that uses AI to limit exposure to harmful or misleading content. This system proactively protects users, especially the younger audience, by filtering out misinformation and unhealthy behaviors. While its primary aim is harm avoidance, it also indirectly bolsters wellbeing by fostering a safer, more reliable online space. This not only prevents the spread of negative content but also enhances user engagement through a more positive and informative experience, thereby improving its audience's overall mental and emotional health. Clearly, this strategy of focusing primarily on harm mitigation is the one most commonly employed currently. It should be clear that although the current predominant strategy in AI emphasizes harm mitigation, this approach may not fully realize the potential benefits that a proactive, positive stance can offer. Such a forward-looking approach is more effective in bringing out positive aspects that go beyond what harm mitigation alone can achieve.

After all, if we keep solving today's problems, we keep intact the system causing the problems in the first place. A visionary designer would instead look for opportunities, even if these require the design of a completely new system (Hekkert & van Dijk, 2011).

6.3.6. Recommendation 6: Positive AI is a moving target, not an endpoint

Lastly, it should be emphasized that the pursuit of Positive AI should not be treated as a box to check but as an ongoing process requiring continued reflection and responsiveness to change. However, many companies approach ethical AI superficially, paying lip service to ethics without meaningful commitment (Morley, Kinsey, et al., 2021). This "ethics washing" allows them to feign social responsibility while continuing development without appropriate oversight (Bietti, 2021). To truly achieve ethical AI, companies must have a real commitment supported by a governance model that continually considers AI's impact on affected groups, integrates ethics into research and decisions, and updates policies accordingly (Morley, Elhalal, et al., 2021).

Therefore, designing truly positive AI necessitates accepting that it is an ongoing pursuit, not an endpoint. This is especially true given the

fluid nature of wellbeing —while the overall concept may be stable, its specific facets shift rapidly, particularly at the intervention pace layer where AI systems operate. What promotes wellbeing through AI today may be inadequate or even counterproductive tomorrow as social norms, technologies, and expectations evolve (see Chapter 3). Rather than seeing positive AI as a fixed destination, companies must commit to the continual effort required to adapt systems to support ever-changing, context-sensitive notions of human wellbeing. Consider the 2023 Israel-Hamas war as an uncomfortable example. Since the onset of this conflict, social media platforms like Instagram and *TikTok* have become flooded with related imagery and messages. Exposure to such content evidently impacts wellbeing during times of conflict. Thus, it would be reasonable to implement wellbeing indicators related to this particular type of content. However, we can reasonably assume this media saturation will fade as the conflict stabilizes and people post less on the topic over time—thus diminishing the need for assessing the impact of war-related content on wellbeing.

Additionally, AI alignment should not be seen as static but as an ongoing process. "Aligned" does not even describe an achievable state for sociotechnical systems as complex as those involving the AIs discussed here. These systems have simply too many shifting components and tradeoffs to identify an optimal, final solution. Hence, the cybernetic concept of a *dynamic equilibrium*. Instead, one can envision alignment as an ongoing process, as a conversation, or perhaps as a process of harmonization akin to the fluctuations seen in typical prey-predator population graphs (see Figure 6.2). In these ecosystems, the populations continuously adjust in a balanced, harmonious pattern rather than ever achieving a static state of harmony—they are harmonizing just as AI systems are aligning. The concept of harmony in design is further discussed in J. D. Lomas and Xue (2022). The idea that wellbeing alignment is no end state resonates with the maximalist conception of alignment, which aims to comprehensively align AI on society-wide or global scales (Gabriel, 2020). This approach implies an ongoing, dynamic process integrating a broader spectrum of ethical, societal, and cultural values. This delineation underscores the key difference between viewing AI alignment as a finite endpoint (the minimalist view) and as a continuous, evolving process (the maximalist perspective).

Furthermore, technological capabilities and sociotechnical contexts continuously evolve. Technology has immense power to reshape society—take the disruptive impact of the *iPhone*, for example. Such innovations introduce entirely new sociotechnical contexts. Namely, the iPhone revolutionized communication and how we access information, engage in commerce, and

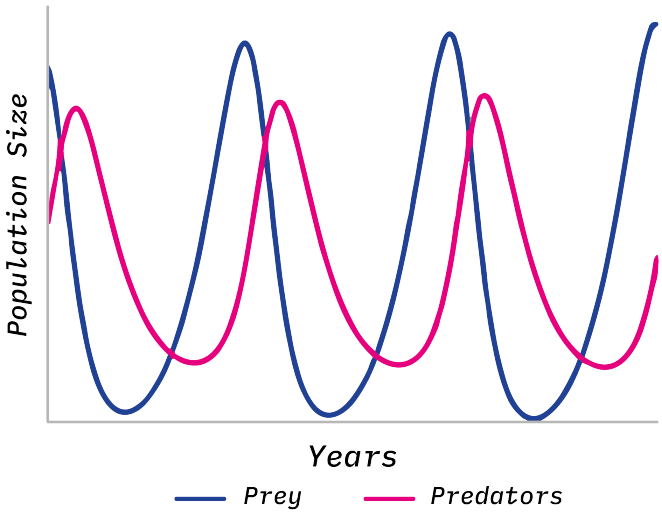


Figure 6.2: A graph showing the harmonious relation between predator and prey populations in an ecosystem.

even perceive and interact with our environment. This innovation led to a significant shift in societal behaviors and norms. In the context of Positive AI, being sensitive to how technological innovation is shaping the ecosystems around us allows us to anticipate and mitigate potential negative impacts while maximizing benefits. Consequently, embracing a maximalist perspective enhances sensitivity to dynamic ecosystems. This approach enables systems to adapt effectively, supporting wellbeing amidst constant technological and social changes.

Finally, we have discussed how interventions in complex sociotechnical systems often result in unintended consequences alongside intended benefits. Additionally, such interventions involve inevitable tradeoffs, making the optimization process inherently challenging. In that vein, Chapter 3 discussed how the pace of change in AI and wellbeing introduces further attribution challenges. While these issues fundamentally make the process of Positive AI difficult and initially costly, refusing to continuously realign risks miscalibrating AI to the wellbeing needs of communities in context. Here, we can return to the principle of satisficing once more: it is not about achieving a comprehensive resolution of optimization tradeoffs instantly but rather about methodically navigating towards this objective. This entails taking deliberate, yet modest steps to address the elements that complicate the equilibrium of tradeoffs methodically. Such an approach helps in clearly

linking interventions to wellbeing outcomes. Tightening this assessment-action loop demonstrates the necessity of persistent realignment to avoid negative repercussions, as observed in Chapter 4's case study. As iterations of My Wellness Check progressed, precision (i.e., our contextual understanding of wellbeing) improved while the need for intervention diminished (i.e., over time, many of the evident wellbeing needs related to COVID-19 were addressed).

Pursuing Positive AI should be treated as a continuous process rather than a static endpoint. As discussed, wellbeing is fluid across contexts, alignment involves inevitable tradeoffs, and sociotechnical systems continuously evolve. Consequently, AI developers must persistently reevaluate AI's impacts and adaptively modify systems to align with dynamic user needs. Rather than superficial ethics washing, achieving lasting positive societal influence necessitates an authentic commitment to iterative realignment. This step-by-step approach enables gradual progress in developing AI that promotes human wellbeing, while allowing for regular reassessment as we work toward beneficial AI.

6.4. Further considerations

This section explores several additional topics relevant to Positive AI. First, connections are drawn between the proposed Positive AI method and the broader field of ethical AI, including its relevance for future advanced systems. Then, reflections are provided on the use of wellbeing as the orienting metric, including potential limitations of this approach. Next, other limitations of the research are acknowledged, including the lack of implementations in industry contexts. Finally, ethical considerations such as privacy concerns and uncritical compliance are discussed.

6.4.1. Future steps: connecting the method to the broader field of AI

As highlighted in Chapter 5, the Positive AI method offers a complementary perspective to enrich existing approaches to develop Positive AI. The next steps to advance the method include connecting to existing alignment methods, empirical validation, expanding case studies, and scaling and generalization

Connecting to other methods. As highlighted in Chapter 5, the Positive AI method offers a complementary perspective to enrich existing approaches to develop ethical AI. For instance, prominent alignment methods focus primarily on technical solutions like reward modeling and algorithm controls.

Meanwhile, popular ethical AI guidelines remain conceptual without concrete design guidance. In contrast, Positive AI has uniquely bridged practical and abstract philosophical principles with participatory, creative design practices centered on wellbeing. It tangibly translates abstract ideals into contextualized measurement and assessment interwoven throughout the design process. So, while respecting the merits of prevailing formal and principled approaches, Positive AI meaningfully integrates rigorous wellbeing considerations directly into iterative human-centered design flows tailored to AI. Moreover, as explained previously, the method could aid the development of benchmarks or aid in the refinement of Constitutional values for RLHF processes (Bai et al., 2022). Further avenues for how academics may support the positive development of AI systems like this should be further explored.

Empirical Validation. The Positive AI method, initially evaluated through student courses and master graduation projects as noted in Chapter 5, could benefit from further validation in more advanced practical settings involving experienced design teams. While collaboration with large platform developers presents challenges, exploring partnerships with smaller firms or engaging in open AI projects like *EleutherAI* offers a viable alternative. Such collaborations could provide a diverse and rich environment for testing and refining the method in real-world applications, enhancing its practicality and relevance in the evolving AI design and development landscape.

Expanding Case Studies. The scarcity of exemplary case studies in Positive AI highlights a need for expansion beyond industry examples. Student courses are a valuable avenue for expanding case studies, offering opportunities to generate new examples and refine the method, particularly in communicating Positive AI's complexities and educational elements. Additionally, considering other contexts for these studies could further enrich the diversity and applicability of Positive AI examples, offering broader insights and practical applications in various domains.

Scaling and Generalization. For companies, the method may become more relevant when scaled up. As emphasized, one of the procedures the method proposes is to be able to translate small-scale research into large-scale optimization metrics. However, this process can be further refined to develop robust ways of doing this that generalize across other domains and contexts.

In conclusion, further progress in the Positive AI agenda could be achieved by first increasing advocacy efforts to raise awareness about the Positive AI approach in both academic and industry circles. Reiterating a point made by (Morley, Elhalal, et al., 2021), a big step forward is greater awareness,

advocacy, and education about ethical AI efforts. Enhancing AI literacy can motivate the broader adoption of ethical principles, but it also empowers the broader public to make more positive use of available technologies. Second, formulating policy recommendations and governance models that support implementing Positive AI principles at organizational and societal levels. Such policies and governance structures can incentivize the integration of wellbeing considerations in AI development. And, third, creating a collaborative community of practice around Positive AI to foster knowledge sharing and innovation. Building connections between researchers, designers, developers, and other stakeholders facilitates continued progress, allowing complementary strengths to synergize around the shared goal of beneficial AI aligned with human flourishing. Through multifaceted efforts spanning advocacy, policy, and collaborative community building, the Positive AI agenda can progressively influence AI innovation trajectories towards the enhancement of societal wellbeing.

6.4.2. Reflecting on wellbeing as orienting metric

This dissertation adopted human wellbeing as the core orienting metric and value for designing positive AI systems. As discussed in the introduction, this perspective aligns with Sam Harris' view that wellbeing provides the ultimate basis for values (S. Harris, 2010). He argues that values like justice and autonomy derive importance from their impact on conscious experience. While certainly debatable, this viewpoint suggests optimizing for wellbeing will inherently account for all values that empirically contribute to human flourishing, ultimately benefiting society at large (Stray, 2020). The case study discussed in Chapter 4 indeed showed that by contextualizing wellbeing, values surfaced that were demonstrably contributive to flourishing in that specific time (COVID-19) and context.

However, one may critique that a singular focus on wellbeing risks oversimplifying the complexity of human values and goals. It may fail to fully capture diverse community interests and lead to unintended consequences from conflicts across values. Inflecting this perspective slightly, S. Harris (2010) would likely admit that a singular focus on wellbeing risks glossing over the full complexity of human values. However, he would argue that all humans share a basic, observable drive to avoid suffering and pursue happiness and fulfillment. While specific concepts of wellbeing clearly differ across individuals and cultures, wellbeing can still serve as an umbrella proxy for evaluation based on common ground. This dissertation recognizes that the initial models that come up in the contextualization phase are too simplistic. That is why reevaluating and scrutinizing these models is

emphasized. As seen in practice in Chapter 4, iterations took place, allowing the model to become more nuanced over time. This nuance is reflected in the changing contents of the wellbeing instrument as topics and questions came and went. The model particularly improved when groups previously not included, such as people living with disabilities, were explicitly asked for their opinions. Thus, while a singular focus on wellbeing risks oversimplifying the complexity of human values and goals, the iterative process allows for the incorporation of diverse perspectives. This enables a more comprehensive consideration of community interests, leading to the evolution of the models to better capture the full picture. The goal is to mitigate unintended consequences from conflicts across values.

Another critique may be the position that there is a universally shared basis for much of a positivist regarding something as complex and "human" as wellbeing. Specifically, is there truly a shared essence of wellbeing that needs uncovering? And why would identifying such an essence, if it exists, be relevant for AI systems aiming to support wellbeing? After all, as discussed in Chapter 3, any given system will only ever capture a partial glimpse of its users' wellbeing. In response, this dissertation balances positivist assumptions with constructivist sensibilities—i.e., a cybernetic framework naturally accommodates and integrates both stances (Yolles, 2021). This integration acknowledges the usefulness of identifying patterns in wellbeing experiences to inform AI optimization while recognizing the diversity of equally valid perspectives. Moreover, the approach seems to align with Sam Harris' view on deriving knowledge of wellbeing. He regards wellbeing as pertinent only when it affects the wellbeing of conscious creatures (S. Harris, 2010). Therefore, while there may be universal truths about wellbeing, our understanding of it is derived from the experiences of these conscious entities (Drob, 2016), necessitating a constructivist epistemology.

Lastly, adopting wellbeing as the primary value risks an anthropocentric perspective that ignores the intrinsic worth of nonhuman entities. Posthumanist thinkers, for example, caution against human-centered worldviews that position humanity as the supreme concern (e.g., Braidotti, 2023). That is, a sole focus on human wellbeing fails to account for animal welfare, ecological sustainability, and the interests of potential future forms of intelligence (Madianou, 2021). This critique challenges the notion of wellbeing as the definitive orienting metric, arguing it promotes an exploitative, hierarchical relationship between humans and our environment. From a posthumanist viewpoint, designing AI systems focused strictly on maximizing human wellbeing could lead to unintended consequences, ignoring interconnected ecological and ethical systems.

However, it should be clear that in this dissertation, the concepts of both human wellbeing and human-centered design are intended to be inclusive of environmental and animal welfare concerns. Improving wellbeing at the cost of the planet or animals would not align with the analogy of desirable routes through the moral landscape (e.g., it risks creating more suffering (S. Harris, 2010). Yet we live in a world still dominated by perspectives that fail to account for these concerns. Therefore, even as we put wellbeing at the helm of our innovations, we must operate within these confines while also broadening our considerations where feasible to include nonhuman perspectives.

6.4.3. Limitations

This research significantly contributes to developing frameworks and methods for designing AI systems focused on human wellbeing. However, it's important to acknowledge its inherent limitations and the need for further inquiry.

One notable aspect is the absence of direct collaboration with industry organizations. The validation of the method, grounded in case studies and student projects, would benefit from actual industry partnerships to reimagine and implement AI systems in real-world scenarios. The absence of collaboration with industry partners is, however, something that plagues the field at large (Stray & Hadfield, n.d.; Stray et al., 2023). Companies are hesitant to allow third-party research to be conducted on their platforms for obvious reasons discussed in Chapter 3. Despite this, the proposed method is designed to be adaptable to various contexts. Therefore, testing the method through applications in professional design settings remains an important opportunity for future work.

Furthermore, integrating global and local cultural values into the wellbeing frameworks developed merits more extensive validation. While the study engaged an international participant group, the research context remained within an academic institution in the Global North. The inclusion of diverse cultural voices, as shown in Chapter 4, provided initial steps towards broadening conceptualizations. However, substantially more deliberate efforts are needed to comprehend perspectives on wellbeing that extend to other contexts. Testing applications of the method across various cultural environments could highlight valuable ways to strengthen the incorporation of multiple worldviews. Additionally, feedback from the method evaluation in Chapter 5 indicates strong interest among designers for tools granting deeper, multicultural understandings of wellbeing. This suggests fertile opportunities to enhance the method's sensitivity to diverse paradigms through further research across geographic and social contexts.

On a different note, the research emphasizes participatory approaches sensitive to contextual nuances but falls short in offering detailed guidance for managing equitable stakeholder involvement and power dynamics. The case study conducted in the context of the COVID-19 pandemic, as detailed in Chapter 4, highlights the facilitation of stakeholder engagement under crisis conditions. However, this crisis situation with a shared wellbeing objective likely does not reflect typical participation barriers or power imbalances arising in corporate settings. Addressing these challenges is essential, particularly in scenarios lacking a unifying crisis. The research offers fertile ground for developing more refined techniques prioritizing diversity, empowerment, and subtle power dynamics in various contexts.

Lastly, the dissertation provides limited discussion regarding potential unintended consequences and downsides associated with the proposed focus on continuous wellbeing measurement and aligned interventions. In particular, while the method in Chapter 5 discusses some aspects of how to proactively consider the consequences of your designs, there is a need for further reflection. It is crucial to develop proactive strategies to mitigate these concerns. These ethical considerations, including the risks of over-reliance and manipulation, among other pitfalls, will be explored further in the following section.

Ethical tradeoffs

While driven by benevolent intentions, using AI systems to optimize wellbeing poses ethical risks that warrant thoughtful mitigation. For instance, people may feel pressured to disclose sensitive personal details in return for optimized interactions. However, an over-reliance on technology has already normalized excessive data collection, often without meaningful consent (Yeung, 2018). As AI recommendations depend on more intimate user insights, ensuring ethical data-sharing practices is paramount to preventing further privacy infringement.

Additionally, dependence on AI directives risks fostering reliance and eroding self-regulatory capacities over time. Much like navigation applications have deskilled innate spatial abilities (Hebblewhite & Gillett, 2021), AI support for wellbeing optimization may inadvertently hinder emotional resilience and coping skills when used as a technological crutch versus a complement to human abilities. Therefore, establishing safeguards, such as nurturing AI literacy skills, is vital to confirm systems augment rather than replace human strengths. Thoughtfully designed interventions should empower sustainable self-driven wellbeing versus long-term dependency on technology. Also, because of the lack of working with companies,

the commercial challenges have not received adequate investigation. In Chapter 1, I introduced that this work focuses on challenges around actively integrating wellbeing, which leads to conflicts with commercial incentives. These challenges of power have remained underexplored (as discussed in Chapter 3). Nonetheless, the fundamental challenges I have addressed will also always apply in cases of active integration—there will just be additional difficulties.

Furthermore, AI-guided recommendations aimed at wellbeing promotion could undermine people's independence if directives override individual agency in an effort to "optimize" outcomes. With social media, users already alter their behavior in hopes of gaining external validation, rather than acting upon intrinsic motivations (Stsiampkouskaya, Joinson, Piwek, & Stevens, 2021). Preventing further manipulation will necessitate ingraining transparency, (scalable) oversight, and integrity principles within these systems to confirm they reinforce human flourishing in alignment with user values versus skewing behavior simply to satisfy algorithms.

Thus, while AI focused on elevating the human experience is undoubtedly well-intentioned, the risks of misuse, over-reliance, and manipulation must be mitigated proactively rather than after the fact. Therefore, resources should be invested in proactive mitigation strategies such as educating both AI developers and end-users (D. T. K. Ng, Leung, Chu, & Qiao, 2021).

6.5. Conclusion

This dissertation has discussed the design of Positive AI, which is defined as AI that actively promotes wellbeing. It highlights the significant role of human-centered designers in guiding AI towards positively impacting human lives. A new methodology is introduced to integrate wellbeing principles into AI design, validated through various case studies and expert insights. This seeks to support AI development by providing methods and framework to align technological innovation with human-centric values.

The research underscores the impact of AI on human experiences and advocates for systems that actively support human flourishing. It stresses the importance of feedback loops between human experience and AI system development mediated by human-centered design and research. The sociotechnical perspective enables designers to look beyond the algorithm and creates new affordances for creating positive social impact in the Age of AI. It empowers anyone to design interactions to make positive use of AI in their own context. This work acts as a guide for future AI endeavors at the intersection of technology and human flourishing, promoting a more human-centered approach to AI development.

Bibliography

Abaza, A. (2021). *Mlops: Why is it the most important technology in the age of ai?* Retrieved from https://www.synapse-analytics.io/post/mlops-why-is-it-the-most-important-technology-in-the-age-of-ai

Abrahams, S., & Balkin, R. S. (2006). Review of the five factor wellness inventory (5f-wel). *News Notes*, *46*(2), 1–3.

Alexandrova, A. (2012, December). Well-Being as an Object of Science. *Philosophy of Science*, *79*(5), 678–689. Retrieved 2022-05-17, from https://www.cambridge.org/core/product/identifier/S0031824800014793/type/journal_article doi: 10.1086/667870

Alexandrova, A. (2017). *A philosophy for the science of well-being.* Oxford University Press.

Alexandrova, A., & Fabian, M. (2022). Democratising measurement: Or why thick concepts call for coproduction. *European Journal for Philosophy of Science*, *12*(1), 7.

Alfrink, K., Keller, I., Kortuem, G., & Doorn, N. (2022, August). Contestable AI by Design: Towards a Framework. *Minds and Machines*, 1–27. Retrieved 2022-08-23, from https://doi.org/10.1007/s11023-022-09611-z doi: 10.1007/s11023-022-09611-z

Allas, T., Chinn, D., Sjatil, P. E., & Zimmerman, W. (2020). Well-being in Europe : Addressing the high cost of COVID-19 on life satisfaction. (June). Retrieved from https://www.mckinsey.com/industries/public-sector/our-insights/safeguarding-europes-livelihoods-mitigating-the-employment-impact-of-covid-19?cid=other-eml-alt-mip-mck&hlkid=a5c37686511e4f48a90f14ff725b5aac&hctky=11221497&hdpid=e0c4935b-989a-491e-b3f6-4608

Andreasen, M. M. (2003). Improving design methods' usability by a mindset approach. *Human Behaviour in Design: Individuals, Teams, Tools*, 209–218.

Anthropic, & Collective Intelligence Project. (2023, 10 17). *Collective constitutional ai: Aligning a language model with public input.* Retrieved from https://www.anthropic.com/collective-constitutional-ai (Retrieved 2024-03-28)

Arnold, R. D., & Wade, J. P. (2015, January). A Definition of Systems Thinking: A Systems Approach. *Procedia Computer Science*, *44*, 669–678. Retrieved 2022-11-09, from https://www.sciencedirect.com/science/article/pii/S1877050915002860 doi: 10.1016/j.procs.2015.03.050

Arora, S., & Doshi, P. (2021). A survey of inverse reinforcement learning: Challenges, methods and progress. *Artificial Intelligence*, *297*, 103500. doi: https://doi.org/10.1016/j.artint.2021.103500

Atkinson, S., Bagnall, A.-M., Corcoran, R., South, J., & Curtis, S. (2020, June). Being Well Together: Individual Subjective and Community Wellbeing. *Journal of Happiness Studies*, *21*(5), 1903–1921. Retrieved 2022-11-16, from http://link.springer.com/10.1007/s10902-019-00146-2 doi: 10.1007/s10902-019-00146-2

Aucejo, E. M., French, J., Paola, M., Araya, U., & Zafar, B. (2020). The impact of COVID-19 on student experiences and expectations : Evidence from a survey. *Journal of Public Economics*, *191*, 104271–104271. Retrieved from https://doi.org/10.1016/j.jpubeco.2020.104271 (Publisher: Elsevier B.V.) doi: 10.1016/j.jpubeco.2020.104271

Auernhammer, J. (2022, August). Human-centered AI: The role of Human-centered Design Research in the development of AI. In *Boess, S., Cheung, M. and Cain, R. (eds.), Synergy - DRS International Conference 2020.* Online. Retrieved 2023-03-28, from https://dl.designresearchsociety.org/drs-conference-papers/drs2020/researchpapers/89 doi: 10.21606/drs.2020.282

Aydin, d., & Karaarslan, E. (2023, September). Is ChatGPT Leading Generative AI? What is Beyond Expectations? *Academic Platform Journal of Engineering and Smart Systems*, *11*(3), 118–134. Retrieved 2023-10-09, from http://dergipark.org.tr/en/doi/10.21541/apjess.1293702 doi: 10.21541/apjess.1293702

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., … Kaplan, J. (2022, December). *Constitutional AI: Harmlessness from AI Feedback.* arXiv. Retrieved 2023-03-22, from http://arxiv.org/abs/2212.08073 (arXiv:2212.08073 [cs])

Baidoo-Anu, D., & Owusu Ansah, L. (2023). Education in the Era of Generative Artificial Intelligence (AI): Understanding the Potential Benefits of ChatGPT in Promoting Teaching and Learning. *SSRN Electronic Journal*. Retrieved 2023-10-09, from https://www.ssrn.com/abstract=4337484 doi: 10.2139/ssrn.4337484

Bakas, T., McLennon, S. M., Carpenter, J. S., Buelow, J. M., Otte, J. L., Hanna, K. M., … Welch, J. L. (2012, November). Systematic review

of health-related quality of life models. *Health and Quality of Life Outcomes*, *10*(1), 134. Retrieved 2022-12-12, from https://doi.org/10.1186/1477-7525-10-134 doi: 10.1186/1477-7525-10-134

Baumgartner, J., Ruettgers, N., Hasler, A., Sonderegger, A., & Sauer, J. (2021). Questionnaire experience and the hybrid System Usability Scale: Using a novel concept to evaluate a new instrument. *International Journal of Human Computer Studies*, *147*, 102575–102575. Retrieved from https://doi.org/10.1016/j.ijhcs.2020.102575 (Publisher: Elsevier Ltd) doi: 10.1016/j.ijhcs.2020.102575

Beardow, C., van der Maden, W., & Lomas, J. (2020). Designing smart system: Reframing artificial intelligence for human-centered designers. , 11–15.

Beer, R. D., Chiel, H. J., & Sterling, L. S. (1990). A biological perspective on autonomous agent design. *Robotics and Autonomous systems*, *6*(1-2), 169–186.

Bentvelzen, M., Woźniak, P. W., Herbes, P. S., Stefanidi, E., & Niess, J. (2022, March). Revisiting Reflection in HCI: Four Design Resources for Technologies that Support Reflection. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, *6*(1), 1–27. Retrieved 2023-04-05, from https://dl.acm.org/doi/10.1145/3517233 doi: 10.1145/3517233

Bietti, E. (2021). From Ethics Washing to Ethics Bashing: A View on Tech Ethics from Within Moral Philosophy. *SSRN Electronic Journal*. Retrieved 2022-05-13, from https://www.ssrn.com/abstract=3914119 doi: 10.2139/ssrn.3914119

Biswas-Diener, R. (2022, October). Wellbeing research needs more cultural approaches. *International Journal of Wellbeing*, *12*(4). Retrieved 2023-10-12, from https://www.internationaljournalofwellbeing.org/index.php/ijow/article/view/1965 (Number: 4) doi: 10.5502/ijw.v12i4.1965

Bleecker, J. (2022). Design fiction: A short essay on design, science, fact, and fiction. In S. Carta (Ed.), *Machine learning and the city: Applications in architecture and urban design* (pp. 463–468). Hoboken, NJ: Wiley. doi: 10.1002/9781119815075.ch47

Blessing, & Chakrabarti. (2009). *Drm, a design research methodology*. Springer Dordrecht.

Blijlevens, J., Thurgood, C., Hekkert, P., Chen, L.-L., Leder, H., & Whitfield, T. (2017). The aesthetic pleasure in design scale: The development of a scale to measure aesthetic pleasure for designed artifacts. *Psychology of Aesthetics, Creativity, and the Arts*, *11*(1), 86.

Bloom, B. S. (1973, March). Recent developments in mastery learning. *Educational Psychologist*, *10*(2), 53–57. Retrieved 2022-08-04, from https://doi.org/10.1080/00461527309529091 (Publisher: Routledge _eprint: https://doi.org/10.1080/00461527309529091) doi: 10.1080/00461527309529091

Boateng, G. O., Neilands, T. B., Frongillo, E. A., Melgar-Quiñonez, H. R., & Young, S. L. (2018). Best Practices for Developing and Validating Scales for Health, Social, and Behavioral Research: A Primer. *Frontiers in Public Health*, *6*(June), 1–18. doi: 10.3389/fpubh.2018.00149

Bono, G., Reil, K., & Hescox, J. (2020, August). Stress and wellbeing in urban college students in the U.S. during the COVID-19 pandemic: Can grit and gratitude help? *International Journal of Wellbeing*, *10*(3). Retrieved 2022-06-01, from https://www.internationaljournalofwellbeing.org/index.php/ijow/article/view/1331 (Number: 3)

Boschman, J. S., Nieuwenhuijsen, K., & Sluiter, J. K. (2018). Within-person fluctuations in wellbeing and task-specific work ability. *Quality of Life Research*, *27*, 437–446. doi: https://doi.org/10.1007/s11136-017-1713-3

Bostrom, N. (2003). Ethical issues in advanced artificial intelligence. *Science fiction and philosophy: from time travel to superintelligence*, 277–284.

Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press. (Google-Books-ID: 7_H8AwAAQBAJ)

Bostrom, N. (2019). The Vulnerable World Hypothesis. *Global Policy*, *10*(4), 455–476. doi: 10.1111/1758-5899.12718

Boy, G. A. (2017). Human-centered design of complex systems: An experience-based approach. *Design Science*, *3*, e8.

Braidotti, R. (2023). Posthuman feminism.

Brand, S. (2018, January). Pace Layering: How Complex Systems Learn and Keep Learning. *Journal of Design and Science*. Retrieved 2022-07-12, from https://jods.mitpress.mit.edu/pub/issue3-brand/release/2 doi: 10.21428/7f2e5f08

Brodeur, A., Clark, A. E., Fleche, S., & Powdthavee, N. (2021). COVID-19 , lockdowns and well-being : Evidence from Google Trends. *Journal of Public Economics*, *193*, 104346–104346. Retrieved from https://doi.org/10.1016/j.jpubeco.2020.104346 (Publisher: Elsevier B.V.) doi: 10.1016/j.jpubeco.2020.104346

Bronfenbrenner, U., et al. (1994). Ecological models of human development. *International encyclopedia of education*, *3*(2), 37–43.

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., … others (2023). Sparks of artificial general intelligence: Early

experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

Burnell, R., Schellaert, W., Burden, J., Ullman, T. D., Martinez-Plumed, F., Tenenbaum, J. B., … others (2023). Rethink reporting of evaluation results in ai. *Science*, *380*(6641), 136–138.

Burns, D., Dagnall, N., & Holt, M. (2020). Assessing the Impact of the COVID-19 Pandemic on Student Wellbeing at Universities in the United Kingdom: A Conceptual Analysis. *Frontiers in Education*, *5*. Retrieved 2022-06-01, from https://www.frontiersin.org/article/10.3389/feduc.2020.582882

Bursztyn, L., Egorov, G., Enikolopov, R., & Petrova, M. (2019). *Social media and xenophobia: Evidence from russia* (Tech. Rep.). National Bureau of Economic Research.

Butler, J., & Kern, M. L. (2016). The PERMA-Profiler: A brief multidimensional measure of flourishing. *International Journal of Wellbeing*, *6*(3), 1–48. doi: 10.5502/ijw.v6i3.526

Callahan, D. (1973). The WHO Definition of 'Health'. *The Hastings Center Studies*, *1*(3), 77–87. Retrieved 2022-12-09, from https://www.jstor.org/stable/3527467 (Publisher: The Hastings Center) doi: 10.2307/3527467

Calvo, R. A., & Peters, D. (2014). *Positive Computing: Technology for Wellbeing and Human Potential*. MIT Press. (Google-Books-ID: uI6ZBQAAQBAJ)

Calvo, R. A., & Peters, D. (2019). Design for wellbeing - Tools for research, practice and ethics. *Conference on Human Factors in Computing Systems - Proceedings*, 1–5. doi: 10.1145/3290607.3298800

Calvo, R. A., Peters, D., & Cave, S. (2020). Advancing impact assessment for intelligent systems. *Nature Machine Intelligence*, *2*(2), 89–91. Retrieved from http://dx.doi.org/10.1038/s42256-020-0151-z (Publisher: Springer US) doi: 10.1038/s42256-020-0151-z

Calvo, R. A., Peters, D., Vold, K., & Ryan, R. M. (2020). Supporting human autonomy in ai systems: A framework for ethical enquiry. *Ethics of digital well-being: A multidisciplinary approach*, 31–54.

Calvo, R. A., Vella-Brodrick, D., Desmet, P., & M. Ryan, R. (2016). Editorial for "Positive Computing: A New Partnership Between Psychology, Social Sciences and Technologists". *Psychology of Well-Being*, *6*(1), 4–9. (Publisher: Springer Berlin Heidelberg) doi: 10.1186/s13612-016-0047-1

Camfield, L. (2016). Enquiries into Wellbeing: How Could Qualitative Data Be Used to Improve the Reliability of Survey Data? In S. C. White & C. Blackmore (Eds.), *Cultures of Wellbeing: Method, Place, Policy*

(pp. 47–65). London: Palgrave Macmillan UK. Retrieved 2023-04-03, from https://doi.org/10.1057/9781137536457_2 doi: 10.1057/9781137536457_2

Cash, P., Daalhuizen, J., & Hekkert, P. (2023). Evaluating the efficacy and effectiveness of design methods: A systematic review and assessment framework. *Design Studies*, *88*, 101204.

Cavalcante Siebert, L., Lupetti, M. L., Aizenberg, E., Beckers, N., Zgonnikov, A., Veluwenkamp, H., … others (2023). Meaningful human control: actionable properties for ai system development. *AI and Ethics*, *3*(1), 241–255.

Cave, S., & Dihal, K. (2020, December). The Whiteness of AI. *Philosophy & Technology*, *33*(4), 685–703. Retrieved 2023-10-12, from https://doi.org/10.1007/s13347-020-00415-6 doi: 10.1007/s13347-020-00415-6

Cecchinato, M. E., Rooksby, J., Hiniker, A., Munson, S., Lukoff, K., Ciolfi, L., … Harrison, D. (2019). Designing for digital wellbeing: A research & practice agenda. In *Extended abstracts of the 2019 chi conference on human factors in computing systems* (pp. 1–8).

Chan, L., Swain, V. D., Kelley, C., de Barbaro, K., Abowd, G. D., & Wilcox, L. (2018). Students' Experiences with Ecological Momentary Assessment Tools to Report on Emotional Well-being. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, *2*(1), 1–20. doi: 10.1145/3191735

Chordia, I., Tran, L.-P., Tayebi, T. J., Parrish, E., Erete, S., Yip, J., & Hiniker, A. (2023). Deceptive design patterns in safety technologies: A case study of the citizen app. In *Proceedings of the 2023 chi conference on human factors in computing systems* (pp. 1–18).

Christian, B. (2020). *The Alignment Problem: Machine Learning and Human Values.* WW Norton. (Google-Books-ID: KGCNEAAAQBAJ)

Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., & Amodei, D. (2017). Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, *30*.

Chui, M., Hazan, E., Roberts, R., Singla, A., & Smaje, K. (2023). The economic potential of generative ai.

Clark, A., & Chalmers, D. (1998). The extended mind. *analysis*, *58*(1), 7–19.

Clifton, J., & Harter, J. (2021). *Wellbeing at Work.* Simon and Schuster. (Google-Books-ID: T5YfEAAAQBAJ)

Commission, E. (2019). *A definition of Artificial Intelligence: main capabilities and scientific disciplines | Shaping Europe's digital future.* Retrieved 2022-11-09, from https://digital-strategy.ec.europa.eu/en/

library/definition-artificial-intelligence-main-capabilities-and-scientific-disciplines

Cooke, P. J., Melchert, T. P., & Connor, K. (2016). Measuring Well-Being: A Review of Instruments. *Counseling Psychologist*, *44*(5), 730–757. doi: 10.1177/0011000016633507

Corbin, C. B., & Pangrazi, R. P. (2001, December). *Toward a Uniform Definition of Wellness: A Commentary* (Tech. Rep.). President's Council on Physical Fitness and Sports, 200 Independence Avenue, S. Retrieved 2022-12-12, from https://eric.ed.gov/?id=ED470691 (Publication Title: President's Council on Physical Fitness and Sports Research Digest ERIC Number: ED470691)

Costanza-Chock, S. (2018, July). Design Justice, A.I., and Escape from the Matrix of Domination. *Journal of Design and Science*. Retrieved 2023-10-12, from https://jods.mitpress.mit.edu/pub/costanza-chock doi: 10.21428/96c8d426

Costanza-Chock, S. (2020). *Design Justice: Community-Led Practices to Build the Worlds We Need*.

Council, D. (2007). Eleven lessons: managing design in eleven global brands. a study of the design process. *Design Council, London, Desk research report*.

Coyne, S. M., Rogers, A. A., Zurcher, J. D., Stockdale, L., & Booth, M. (2020). Does time spent using social media impact mental health?: An eight year longitudinal study. *Computers in Human Behavior*, *104*(July 2019), 106160–106160. Retrieved from https://doi.org/10.1016/j.chb.2019.106160 (Publisher: Elsevier Ltd) doi: 10.1016/j.chb.2019.106160

Crawford, D. N. (2020). Supporting student wellbeing during covid-19: Tips from regional and remote australia.

Crawford, K. (2021). *The atlas of ai: Power, politics, and the planetary costs of artificial intelligence*. Yale University Press.

Crawford, K., & Calo, R. (2016). There is a blind spot in ai research. *Nature*, *538*, 311–313. doi: https://doi.org/10.1038/538311a

Cross, N. (2023). *Design thinking: Understanding how designers think and work*. Bloomsbury Publishing.

Csikszentmihalyi, M., Abuhamdeh, S., & Nakamura, J. (2005). Flow. *Handbook of competence and motivation*, 598–608.

Csikszentmihalyi, M., Csikszentmihalyi, M., Abuhamdeh, S., & Nakamura, J. (2014). Flow. *Flow and the foundations of positive psychology: The collected works of Mihaly Csikszentmihalyi*, 227–238.

Cugurullo, F., & Acheampong, R. A. (2023, January). Fear of AI: an inquiry

into the adoption of autonomous cars in spite of fear, and a theoretical framework for the study of artificial intelligence technology acceptance. *AI & SOCIETY*. Retrieved 2023-12-20, from https://doi.org/10.1007/s00146-022-01598-6 doi: 10.1007/s00146-022-01598-6

Cummins, R. A., Mccabe, M. P., Romeo, Y., & Gullone, E. (1994). Validity studies the comprehensive quality of life scale (comqol): Instrument development and psychometric evaluation on college staff and students. *Educational and Psychological Measurement*, *54*(2), 372–382.

Daher, M., Carré, P. D., Jaramillo, A., Olivares, H., & Tomicic, A. (2017). Experience and meaning in qualitative research: A conceptual review and a methodological device proposal. *Forum Qualitative Sozialforschung*, *18*(3). doi: 10.17169/fqs-18.3.2696

Dalpiaz, F., & Parente, M. (2019). Re-swot: from user feedback to requirements via competitor analysis. In *Requirements engineering: Foundation for software quality: 25th international working conference, refsq 2019, essen, germany, march 18–21, 2019, proceedings 25* (pp. 55–70).

Dang, J., King, K. M., & Inzlicht, M. (2020). Why are self-report and behavioral measures weakly correlated? *Trends in cognitive sciences*, *24*(4), 267–269.

Daw, T. M., Coulthard, S., Cheung, W. W. L., Brown, K., Abunge, C., Galafassi, D., … Munyi, L. (2015, June). Evaluating taboo trade-offs in ecosystems services and human well-being. *Proceedings of the National Academy of Sciences*, *112*(22), 6949–6954. Retrieved 2024-01-03, from https://www.pnas.org/doi/abs/10.1073/pnas.1414900112 (Publisher: Proceedings of the National Academy of Sciences) doi: 10.1073/pnas.1414900112

Dean, S., Gilbert, T. K., Lambert, N., & Zick, T. (2021). Axes for sociotechnical inquiry in ai research. *IEEE Transactions on Technology and Society*, *2*(2), 62–70.

Deci, E. L., & Ryan, R. M. (2008a). Hedonia, eudaimonia, and well-being: An introduction. *Journal of Happiness Studies*, *9*(1), 1–11. doi: 10.1007/s10902-006-9018-1

Deci, E. L., & Ryan, R. M. (2008b). Self-determination theory: A macrotheory of human motivation, development, and health. *Canadian Psychology*, *49*(3), 182–185. doi: 10.1037/a0012801

Delle Fave, A., Brdar, I., Freire, T., Vella-Brodrick, D., & Wissing, M. P. (2011). The eudaimonic and hedonic components of happiness: Qualitative and quantitative findings. *Social indicators research*, *100*, 185–207.

De Pue, S., Gillebert, C., Dierckx, E., Vanderhasselt, M.-A., De Raedt, R., & Van den Bussche, E. (2021, February). The impact of the COVID-19 pandemic on wellbeing and cognitive functioning of older adults. *Scientific Reports*, *11*(1), 4636. Retrieved 2022-06-01, from https://www.nature.com/articles/s41598-021-84127-7 (Number: 1 Publisher: Nature Publishing Group) doi: 10.1038/s41598-021-84127-7

Desmet, P. M. A., & Fokkinga, S. (2020). Beyond Maslow ' s Pyramid : Introducing a Typology of Thirteen Fundamental Needs for Human-Centered Design.

Desmet, P. M. A., & Pohlmeyer, A. E. (2013). Positive design: An introduction to design for subjective well-being. *International Journal of Design*, *7*(3), 5–19.

Diener, E. (2019). *Assessing Well-being* (Vol. 53). (Issue: 9 Pages: 1791 Publication Title: Fitzpatrick's Dermatology) doi: 10.1017/CBO9781107415324.004

Diener, E., Emmons, R. A., Larsen, R., & Griffin, S. (1985). The Satisfaction With Life Scale. *Journal of Personality Assessment*, *49*(1), 71–75. doi: 10.23943/princeton/9780691188959.003.0002

Diener, E., & Michalos, A. C. (Eds.). (2009). *Assessing Well-Being* (Vol. 39). Dordrecht: Springer Netherlands. Retrieved 2022-05-04, from http://link.springer.com/10.1007/978-90-481-2354-4 doi: 10.1007/978-90-481-2354-4

Diener, E., & Ryan, K. (2008). Subjective well-being : a general overview. , *39*(4), 391–406.

Diener, E., Suh, E. M., Lucas, R. E., & Smith, H. L. (1999). Subjective well-being: Three decades of progress. *Psychological Bulletin*, *125*(2), 276–302. doi: 10.1037/0033-2909.125.2.276

Diener, E., Wirtz, D., Tov, W., Kim-Prieto, C., Choi, D. w., Oishi, S., & Biswas-Diener, R. (2010). New well-being measures: Short scales to assess flourishing and positive and negative feelings. *Social Indicators Research*, *97*(2), 143–156. doi: 10.1007/s11205-009-9493-y

Dikker, S., Wan, L., Davidesco, I., Kaggen, L., Oostrik, M., McClintock, J., … others (2017). Brain-to-brain synchrony tracks real-world dynamic group interactions in the classroom. *Current biology*, *27*(9), 1375–1380.

Di Vaio, A., Palladino, R., Hassan, R., & Escobar, O. (2020). Artificial intelligence and business models in the sustainable development goals perspective: A systematic literature review. *Journal of Business Research*, *121*, 283–314.

Dobbe, R., Krendl Gilbert, T., & Mintz, Y. (2021, November). Hard choices in artificial intelligence. *Artificial Intelligence*, *300*, 103555.

Retrieved 2023-04-07, from https://linkinghub.elsevier.com/retrieve/pii/S0004370221001065 doi: 10.1016/j.artint.2021.103555

Dodge, R., Daly, A., Huyton, J., & Sanders, L. (2012, August). The challenge of defining wellbeing. *International Journal of Wellbeing*, *2*(3), 222–235. Retrieved 2022-07-25, from http://www.internationaljournalofwellbeing.org/index.php/ijow/article/view/89/238 doi: 10.5502/ijw.v2i3.4

Dolan, P. (2014). *Happiness by design: Change what you do, not how you think*. Penguin.

Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM*, *55*(10), 78–87.

Dorst, K. (2011). The core of 'design thinking'and its application. *Design studies*, *32*(6), 521–532.

Drob, S. L. (2016). An axiological model of the relationship between consciousness and value. *New Ideas in Psychology*, *43*, 57-63. Retrieved from https://www.sciencedirect.com/science/article/pii/S0732118X16000040 doi: https://doi.org/10.1016/j.newideapsych.2016.02.002

Dubberly, H., & Pangaro, P. (2007). Cybernetics and service-craft: Language for behavior-focused design. *Kybernetes*, *36*(9-10), 1301–1317. doi: 10.1108/03684920710827319

Dubberly, H., & Pangaro, P. (2010). *Introduction to Cybernetics and the Design of Systems*. Retrieved from https://www.pangaro.com/CUSO2014/Cybernetics_Book_of_Models-v4.6b-complete.pdf

Dubberly, H., & Pangaro, P. (2019). Cybernetics and Design: Conversations for Action. *Design Research Foundations*, 85–99. doi: 10.1007/978-3-030-18557-2_4

Dung, L. (2023). Current cases of ai misalignment and their implications for future risks. *Synthese*, *202*(5), 138.

D'Alfonso, S. (2020, December). AI in mental health. *Current Opinion in Psychology*, *36*, 112–117. Retrieved 2022-08-04, from https://www.sciencedirect.com/science/article/pii/S2352250X2030049X doi: 10.1016/j.copsyc.2020.04.005

Eason, K. D. (1984). Towards the experimental study of usability. *Behaviour & Information Technology*, *3*(2), 133–143.

Easterlin, R., & Sawangfa, O. (2007). Happiness and Domain Satisfaction: Theory and Evidence. *IZA Discussion Paper*(2584).

Ehsan, U., Liao, Q. V., Muller, M., Riedl, M. O., & Weisz, J. D. (2021). Expanding explainability: Towards social transparency in ai systems. In *Proceedings of the 2021 chi conference on human factors in*

computing systems (pp. 1–19).

Evans, O. G. (2020). *Bronfenbrenner's ecological systems theory.* Simply Psychology. https://www. simplypsychology. org/Bronfenbrenner. html.

Everitt, T., Lea, G., & Hutter, M. (2018). Agi safety literature review. *arXiv preprint arXiv:1805.01109*.

Farzaneh, H. H., & Neuner, L. (2019). Usability evaluation of software tools for engineering design. In *Proceedings of the design society: International conference on engineering design* (Vol. 1, pp. 1303–1312).

Fawaz, M., & Samaha, A. (2021). E-learning: Depression, anxiety, and stress symptomatology among Lebanese university students during COVID-19 quarantine. *Nursing Forum*, *56*(1), 52–57. Retrieved 2022-06-01, from https://onlinelibrary.wiley.com/doi/abs/10.1111/nuf.12521 (_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/nuf.12521) doi: 10.1111/nuf.12521

Felce, D., & Perry, J. (1995, January). Quality of life: Its definition and measurement. *Research in Developmental Disabilities*, *16*(1), 51–74. Retrieved 2022-12-12, from https://www.sciencedirect.com/science/article/pii/0891422294000288 doi: 10.1016/0891-4222(94)00028-8

Felzmann, H., Fosch-Villaronga, E., Lutz, C., & Tamò-Larrieux, A. (2020). Towards transparency by design for artificial intelligence. *Science and Engineering Ethics*, *26*(6), 3333–3361.

Fietta, V., Zecchinato, F., Stasi, B. D., Polato, M., & Monaro, M. (2022, June). Dissociation Between Users' Explicit and Implicit Attitudes Toward Artificial Intelligence: An Experimental Study. *IEEE Transactions on Human-Machine Systems*, *52*(3), 481–489. Retrieved 2023-12-20, from https://ieeexplore.ieee.org/document/9627595 (Conference Name: IEEE Transactions on Human-Machine Systems) doi: 10.1109/THMS.2021.3125280

Fioramonti, L., Coscieme, L., Costanza, R., Kubiszewski, I., Trebeck, K., Wallis, S., … De Vogli, R. (2022). Wellbeing economy: An effective paradigm to mainstream post-growth policies? *Ecological Economics*, *192*(October 2021), 107261–107261. Retrieved from https://doi.org/10.1016/j.ecolecon.2021.107261 (Publisher: Elsevier B.V.) doi: 10.1016/j.ecolecon.2021.107261

Floridi, L. (2023). Ai as agency without intelligence: on chatgpt, large language models, and other generative models. *Philosophy & Technology*, *36*(1), 15.

Floridi, L., & Cowls, J. (2019). A Unified Framework of Five Principles for AI in Society. *Harvard Data Science Review*, 1–15. doi: 10.1162/

99608f92.8cd550d1

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., … Vayena, E. (2018). AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds and Machines*, *28*(4), 689–707. Retrieved from https://doi.org/10.1007/s11023-018-9482-5 (Publisher: Springer Netherlands) doi: 10.1007/s11023-018-9482-5

Fokkinga, S. (2015). *Negative emotions for positive experiences.* Delft: Ipskamp drukkers. (Pages: 254)

Fokkinga, S., Desmet, P., & Hekkert, P. (2020). Impact-centered design: Introducing an integrated framework of the psychological and behavioral effects of design. *International Journal of Design*, *14*(3), 97.

Fordyce, M. W. (1988). A review of research on the happiness measures: A sixty second index of happiness and mental health. *Social indicators research*, *20*, 355–381.

Freitas, C. P. P., Damásio, B. F., Kamei, H. H., Tobo, P. R., Koller, S. H., & Robitschek, C. (2018). Personal growth initiative scale-ii: Adaptation and psychometric properties of the brazilian version. *Paidéia (Ribeirão Preto)*, *28*.

Friedman, B., Kahn, P. H., & Borning, A. (2009). Value Sensitive Design and Information Systems. *The Handbook of Information and Computer Ethics*(May), 69–101. doi: 10.1002/9780470281819.ch4

Frijters, P., & Krekel, C. (2021). *A Handbook for Wellbeing Policy-making: History, Theory, Measurement, Implementation, and Examples.* Oxford University Press. (Google-Books-ID: iekrEAAAQBAJ)

Frisch, M. B., Cornell, J., Villanueva, M., & Retzlaff, P. J. (1992). Clinical validation of the quality of life inventory. a measure of life satisfaction for use in treatment planning and outcome assessment. *Psychological assessment*, *4*(1), 92.

Future of Life Institute. (2023). *Pause Giant AI Experiments: An Open Letter.* Retrieved 2023-12-02, from https://futureoflife.org/open-letter/pause-giant-ai-experiments/

Gabriel, I. (2020). Artificial Intelligence, Values, and Alignment. *Minds and Machines*, *30*(3), 411–437. Retrieved from https://doi.org/10.1007/s11023-020-09539-2 (Publisher: Springer Netherlands) doi: 10.1007/s11023-020-09539-2

Gaggioli, A., Riva, G., Peters, D., & Calvo, R. A. (2017). Positive Technology, Computing, and Design: Shaping a Future in Which Technology Promotes Psychological Well-Being. *Emotions and Affect in Human Factors and Human-Computer Interaction*(April), 477–502.

doi: 10.1016/B978-0-12-801851-4.00018-5

Gates, B. (2023, March). The age of AI has begun. *Gates Notes*. Retrieved 2023-07-25, from https://www.gatesnotes.com/The-Age-of-AI-Has-Begun

Gaver, W. (2012). What should we expect from research through design? In *Proceedings of the sigchi conference on human factors in computing systems* (pp. 937–946).

Genç, E., & Arslan, G. (2021, December). Optimism and dispositional hope to promote college students' subjective well-being in the context of the COVID-19 pandemic. *Journal of Positive School Psychology*, *5*(2), 87–96. Retrieved 2022-06-01, from https://www.journalppw.com/index.php/jpsp/article/view/127 (Number: 2)

Giacomin, J. (2014, December). What Is Human Centred Design? *The Design Journal*, *17*(4), 606–623. Retrieved 2022-05-10, from https://www.tandfonline.com/doi/full/10.2752/175630614X14056185480186 doi: 10.2752/175630614X14056185480186

Gillespie, T. (2014, February). The Relevance of Algorithms. In T. Gillespie, P. J. Boczkowski, & K. A. Foot (Eds.), *Media Technologies* (pp. 167–194). The MIT Press. Retrieved 2022-07-25, from http://mitpress.universitypressscholarship.com/view/10.7551/mitpress/9780262525374.001.0001/upso-9780262525374-chapter-9 doi: 10.7551/mitpress/9780262525374.003.0009

Glanville, R. (2007). Try again. Fail again. Fail better: The cybernetics in design and the design in cybernetics. *Kybernetes*, *36*(9-10), 1173–1206. doi: 10.1108/03684920710827238

Glanville, R. (2009). System Science and Cybernetics. , *III*, 59–86.

Glanville, R. (2014). How design and cybernetics reflect each other. *Proceedings of Third Symposium of Relating Systems Thinking to Design, Oslo, Norway. October*(February 2015), 15–17.

Goodhart, C. (1975). Problems of monetary management: The uk experience. In *Papers in monetary economics* (Vol. I). Reserve Bank of Australia.

Graham, J., Haidt, J., & Nosek, B. A. (2009). Liberals and conservatives rely on different sets of moral foundations. *Journal of personality and social psychology*, *96*(5), 1029.

Graham, S., Depp, C., Lee, E. E., Nebeker, C., Tu, X., Kim, H.-C., & Jeste, D. V. (2019, November). Artificial Intelligence for Mental Health and Mental Illnesses: An Overview. *Current psychiatry reports*, *21*(11), 116. Retrieved 2022-11-09, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7274446/ doi: 10.1007/s11920-019-1094-0

Green, M. C. (2004). Transportation into narrative worlds: The role of prior knowledge and perceived realism. *Discourse processes*, *38*(2), 247–266.

Gregory, T., Engelhardt, D., Lewkowicz, A., Luddy, S., Guhn, M., Gadermann, A., … Brinkman, S. (2019). Validity of the Middle Years Development Instrument for Population Monitoring of Student Wellbeing in Australian School Children. *Child Indicators Research*, *12*(3), 873–899. (Publisher: Child Indicators Research) doi: 10.1007/s12187-018-9562-3

Group, T. W. (1998). The world health organization quality of life assessment (whoqol): development and general psychometric properties. *Social science & medicine*, *46*(12), 1569–1585.

Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., & Yang, G.-Z. (2019). Xai—explainable artificial intelligence. *Science robotics*, *4*(37), eaay7120.

Hadfield-Menell, D., & Hadfield, G. K. (2019, January). Incomplete Contracting and AI Alignment. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 417–422). Honolulu HI USA: ACM. Retrieved 2023-01-11, from https://dl.acm.org/doi/10.1145/3306618.3314250 doi: 10.1145/3306618.3314250

Hagendorff, T. (2020). The ethics of ai ethics: An evaluation of guidelines. *Minds and Machines*, *30*(1), 99–120. Retrieved from https://doi.org/10.1007/s11023-020-09517-8 doi: 10.1007/s11023-020-09517-8

Hagey, K., & Horwitz, J. (2021). Facebook Tried to Make Its Platform a Healthier Place. It Got Angrier Instead. - WSJ. *Wall Street Journal (Online)*, 1–16. Retrieved from https://www.wsj.com/articles/facebook-algorithm-change-zuckerberg-11631654215?mod=searchresults_pos2&page=1

Hailey, J. (2008). Ubuntu: A literature review. *Document. London: Tutu Foundation*.

Halleröd, B., & Seldén, D. (2013). The multi-dimensional characteristics of wellbeing: How different aspects of wellbeing interact and do not interact with each other. *Social Indicators Research*, *113*, 807-825. Retrieved from https://api.semanticscholar.org/CorpusID:143366597

Han, T. A., Pereira, L. M., Lenaerts, T., & Santos, F. C. (2021). Mediating artificial intelligence developments through negative and positive incentives. *PLoS one*, *16*(1), e0244592.

Harari, Y., Harris, T., & Raskin, A. (2023, March). You Can Have the Blue Pill or the Red Pill, and We're Out of Blue Pills. *The New York Times*. Retrieved 2023-09-12, from https://www.nytimes.com/2023/03/24/opinion/yuval-harari-ai-chatgpt.html

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

Harris, S. (2010). *The Moral Landscape: How Science Can Determine Human Values*. Simon and Schuster. (Google-Books-ID: VttdxFt4kT4C)

Harris, T. (2017, Jul). *Technology is downgrading humanity; let's reverse that trend now.* Retrieved from https://www.commerce.senate.gov/services/files/96E3A739-DC8D-45F1-87D7-EC70A368371D

Havrda, M., & Klocek, A. (2023, April). Well-being impact assessment of artificial intelligence - A search for causality and proposal for an open platform for well-being impact assessment of AI systems. *Evaluation and Program Planning*, *99*, 102294. doi: 10.1016/j.evalprogplan.2023.102294

Havrda, M., & Rakova, B. (2020). Enhanced well-being assessment as basis for the practical implementation of ethical and rights-based normative principles for ai. In *2020 ieee international conference on systems, man, and cybernetics (smc)* (pp. 2754–2761).

Haybron, D. M. (2008). *The pursuit of unhappiness: The elusive psychology of well-being*. Oxford University Press, USA.

Hebblewhite, W., & Gillett, A. J. (2021). Every step you take, we'll be watching you: nudging and the ramifications of gps technology. *AI & SOCIETY*, *36*, 863–875.

Hekkert, P., & van Dijk, M. (2011). *Vision in design - A guidebook for innovators*. Amsterdam: BIS Publishers.

Hemphill, T. A. (2019, May). 'Techlash', responsible innovation, and the self-regulatory organization. *Journal of Responsible Innovation*, *6*(2), 240–247. Retrieved 2022-05-04, from https://www.tandfonline.com/doi/full/10.1080/23299460.2019.1602817 doi: 10.1080/23299460.2019.1602817

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., & Steinhardt, J. (2020). Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Hitokoto, H., & Uchida, Y. (2015). Interdependent happiness: Theoretical importance and measurement validity. *Journal of Happiness Studies*, *16*, 211–239.

Hjetland, G. J., Schønning, V., Aasan, B. E. V., Hella, R. T., & Skogen, J. C. (2021, January). Pupils' Use of Social Media and Its Relation to Mental Health from a School Personnel Perspective: A Preliminary Qualitative Study. *International Journal of Environmental*

Research and Public Health, *18*(17), 9163. Retrieved 2023-12-24, from https://www.mdpi.com/1660-4601/18/17/9163 (Number: 17 Publisher: Multidisciplinary Digital Publishing Institute) doi: 10 .3390/ijerph18179163

Horwitz, A. V., & Wakefield, J. C. (2007). *The loss of sadness: How psychiatry transformed normal sorrow into depressive disorder*. Oxford University Press.

Hsu, Y.-C., Verma, H., Mauri, A., Nourbakhsh, I., Bozzon, A., et al. (2022). Empowering local communities using artificial intelligence. *Patterns*, *3*(3).

Hu, C., Chen, C., & Dong, X.-P. (2021). Impact of COVID-19 Pandemic on Patients With Neurodegenerative Diseases. *Frontiers in Aging Neuroscience*, *13*. Retrieved 2022-06-01, from https://www.frontiersin .org/article/10.3389/fnagi.2021.664965

Hu, K. (2023, February). ChatGPT sets record for fastest-growing user base - analyst note. *Reuters*. Retrieved 2023-10-09, from https://www.reuters.com/technology/chatgpt-sets-record-fastest -growing-user-base-analyst-note-2023-02-01/

Huppert, F. A. (2017). Challenges in defining and measuring well-being and their implications for policy. *Future directions in well-being: Education, organizations and policy*, 163–167.

IEEE SA. (2020). *IEEE 7010-2020 - IEEE Recommended Practice for Assessing the Impact of Autonomous and Intelligent Systems on Human Well-Being*. Retrieved from https://standards.ieee.org/standard/7010 -2020.html

Jacobs, N., & Huldtgren, A. (2021). Why value sensitive design needs ethical commitments. *Ethics and Information Technology*, *23*(1), 23–26. Retrieved from https://doi.org/10.1007/s10676-018-9467-3 doi: 10.1007/s10676-018-9467-3

Jakesch, M., Buçinca, Z., Amershi, S., & Olteanu, A. (2022). How different groups prioritize ethical values for responsible ai. In *2022 acm conference on fairness, accountability, and transparency* (pp. 310–323). doi: https://doi.org/10.1145/3531146.3533097

Jaramillo, S. J., Pohlmeyer, A., & Desmet, P. (2015). *Positive design reference guide*. Delft University of Technology.

Ji, J., Qiu, T., Chen, B., Zhang, B., Lou, H., Wang, K., … others (2023). Ai alignment: A comprehensive survey. *arXiv preprint arXiv:2310.19852*.

Johannes, N., Dienlin, T., Bakhshi, H., & Przybylski, A. K. (2022, January). No effect of different types of media on well-being. *Scientific Reports*, *12*(1), 61. Retrieved 2023-01-31, from https://www.nature.com/

articles/s41598-021-03218-7 (Number: 1 Publisher: Nature Publishing Group) doi: 10.1038/s41598-021-03218-7

Joshanloo, M. (2018). Optimal human functioning around the world: A new index of eudaimonic well-being in 166 nations. *British Journal of Psychology*, *109*(4), 637–655.

Joshanloo, M., Van de Vliert, E., & Jose, P. E. (2021). Four fundamental distinctions in conceptions of wellbeing across cultures. In *The palgrave handbook of positive education* (pp. 675–703). Springer International Publishing Cham.

Joshanloo, M., & Weijers, D. (2019). Islamic perspectives on wellbeing. *Positive psychology in the Middle East/North Africa: Research, policy, and practise*, 237–256.

Keleş, B. Y., McCrae, N., & Grealish, A. (2019). A systematic review: the influence of social media on depression, anxiety and psychological distress in adolescents. *International Journal of Adolescence and Youth*, *25*, 79-93. doi: 10.1080/02673843.2019.1590851

Kennedy, J. F. (1963, 10 03). *Remarks in heber springs, arkansas, at the dedication of greers ferry dam.* https://www.presidency.ucsb.edu/documents/remarks-heber-springs-arkansas-the-dedication-greers-ferry-dam. (Retrieved on January 13, 2024)

Kern, M. L., Williams, P., Spong, C., Colla, R., Sharma, K., Downie, A., … Oades, L. G. (2020). Systems informed positive psychology. *The Journal of Positive Psychology*, *15*(6), 705–715.

Keyes, C. L., Shmotkin, D., & Ryff, C. D. (2002). Optimizing well-being: The empirical encounter of two traditions. *Journal of personality and social psychology*, *82*(6), 1007–1022.

Khan, I., Shah, D., & Shah, S. S. (2021, February). COVID-19 pandemic and its positive impacts on environment: an updated review. *International Journal of Environmental Science and Technology*, *18*(2), 521–530. Retrieved 2022-06-01, from https://doi.org/10.1007/s13762-020-03021-3 doi: 10.1007/s13762-020-03021-3

Kissinger, H. A., Schmidt, E., & Huttenlocher, D. (2021). *The age of ai: and our human future.* Hachette UK.

Kitayama, S., & Markus, H. R. (2000). The pursuit of happiness and the realization of sympathy: Cultural patterns of self, social relations, and well-being. *Culture and subjective well-being*, *1*, 113–161.

Kjell, O. N. (2011). Sustainable Well-Being: A Potential Synergy Between Sustainability and Well-Being Research. *Review of General Psychology*, *15*(3), 255–266. doi: 10.1037/a0024603

Kjell, O. N., & Diener, E. (2021). Abbreviated three-item versions of the

satisfaction with life scale and the harmony in life scale yield as strong psychometric properties as the original scales. *Journal of Personality Assessment*, *103*(2), 183–194.

Klenk, M. (2021). How do technological artefacts embody moral values? *Philosophy & Technology*, *34*(3), 525–544.

Klenk, M., & Duijf, H. (2021, November). Ethics of digital contact tracing and COVID-19: who is (not) free to go? *Ethics and Information Technology*, *23*(1), 69–77. Retrieved 2022-06-02, from https://doi.org/10.1007/s10676-020-09544-0 doi: 10.1007/s10676-020-09544-0

Klincewicz, M., Frank, L. E., & Jane, E. (2022). The ethics of matching: Mobile and web-based dating and hook up platforms. In B. D. Earp, C. Chambers, & L. Watson (Eds.), *Routledge handbook of philosophy of sex and sexuality.* Routledge.

König, L. M., Attig, C., Franke, T., & Renner, B. (2021). Barriers to and facilitators for using nutrition apps: systematic review and conceptual framework. *JMIR mHealth and uHealth*, *9*(6), e20037.

Konu, A., & Rimpelä, M. (2002, March). Well-being in schools: a conceptual model. *Health Promotion International*, *17*(1), 79–87. Retrieved 2023-12-23, from https://doi.org/10.1093/heapro/17.1.79 doi: 10.1093/heapro/17.1.79

Kramer, J., Agogino, A. M., & Roschuni, C. (2016, August). Characterizing Competencies for Human-Centered Design. In *Volume 7: 28th International Conference on Design Theory and Methodology* (p. V007T06A026). Charlotte, North Carolina, USA: American Society of Mechanical Engineers. Retrieved 2023-12-22, from https://asmedigitalcollection.asme.org/IDETC-CIE/proceedings/IDETC-CIE2016/50190/Charlotte,%20North%20Carolina,%20USA/258472 doi: 10.1115/DETC2016-60085

Krippendorff, K. (2007). The cybernetics of design and the design of cybernetics. *Kybernetes*, *36*(9-10), 1381–1392. doi: 10.1108/03684920710827364

Krippendorff, K. (2021, October). *From Uncritical Design to Critical Examinations of its Systemic Consequences.* Retrieved 2022-07-14, from https://rsdsymposium.org/professor-dr-klaus-krippendorff/

Krippendorff, K. (2023). A critical cybernetics. *Constructivist Foundations*, *19*(1), 82–93.

Kross, E., Verduyn, P., Sheppes, G., Costello, C. K., Jonides, J., & Ybarra, O. (2021). Social Media and Well-Being: Pitfalls, Progress, and Next Steps. *Trends in Cognitive Sciences*, *25*(1), 55–66. Retrieved from https://doi.org/10.1016/j.tics.2020.10.005 (Publisher: The Authors)

doi: 10.1016/j.tics.2020.10.005

Krys, K., Haas, B. W., Igou, E. R., Kosiarczyk, A., Kocimska-Bortnowska, A., Kwiatkowska, A., … others (2023). Introduction to a culturally sensitive measure of well-being: Combining life satisfaction and interdependent happiness across 49 different cultures. *Journal of Happiness Studies*, *24*(2), 607–627.

Larsson, S., & Heintz, F. (2020). Transparency in artificial intelligence. *Internet Policy Review*, *9*(2).

Layard, R., & De Neve, J.-E. (2023). *Wellbeing.* Cambridge University Press.

Lazarus, R. S. (2003). Does the positive psychology movement have legs? *Psychological inquiry*, *14*(2), 93–109.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *nature*, *521*(7553), 436–444.

Legg, S., & Hutter, M. (2007, June). *A Collection of Definitions of Intelligence.* arXiv. Retrieved 2022-07-12, from http://arxiv.org/abs/0706.3639 (arXiv:0706.3639 [cs])

Li, F., & Lu, Y. (2021). Engaging end users in an ai-enabled smart service design-the application of the smart service blueprint scape (ssbs) framework. *Proceedings of the Design Society*, *1*, 1363–1372.

Liao, Q. V., & Muller, M. (2019). Enabling value sensitive ai systems through participatory design fictions. *arXiv preprint arXiv:1912.07381.*

Linton, M. J., Dieppe, P., & Medina-Lara, A. (2016). Review of 99 self-report measures for assessing well-being in adults: Exploring dimensions of well-being and developments over time. *BMJ Open*, *6*(7). doi: 10.1136/bmjopen-2015-010641

Liscio, E., Lera-Leri, R., Bistaffa, F., Dobbe, R. I., Jonker, C. M., Lopez-Sanchez, M., … Murukannaiah, P. K. (2023). Value inference in sociotechnical systems. In *Proceedings of the 2023 international conference on autonomous agents and multiagent systems* (pp. 1774–1780).

Liscio, E., van der Meer, M., Siebert, L. C., Jonker, C. M., Mouter, N., & Murukannaiah, P. K. (2021). Axies : Identifying and Evaluating Context-Specific Values. *Aamas 2021*, 799–808.

Living in a brave new AI era. (2023, November). *Nature Human Behaviour*, *7*(11), 1799–1799. Retrieved from https://doi.org/10.1038/s41562-023-01775-7 doi: 10.1038/s41562-023-01775-7

Lomas, J. D., Lin, A., Dikker, S., Forster, D., Lupetti, M. L., Huisman, G., … others (2022). Resonance as a design strategy for ai and social robots. *Frontiers in neurorobotics*, *16*, 850489.

Lomas, J. D., Matzat, U., Pei, L., Rouwenhorst, C., van der Maden, W.,

Stevens, T., … Klaassen, R. (2021). The impact of COVID-19 on university teaching and learning: Evidence for the central importance of student and staff well-being.

Lomas, J. D., Patel, N., & Forlizzi, J. (2021). Designing Data-Informed Intelligent Systems to Create Positive Impact. *Relating Systems And Design Thinking Sysmposium 10*.

Lomas, J. D., & Xue, H. (2022). Harmony in design: a synthesis of literature from classical philosophy, the sciences, economics, and design. *She Ji: The Journal of Design, Economics, and Innovation*, *8*(1), 5–64.

Lomas, T., Waters, L., Williams, P., Oades, L. G., & Kern, M. L. (2021). Third wave positive psychology: Broadening towards complexity. *The Journal of Positive Psychology*, *16*(5), 660–674.

London, A. J. (2019). Artificial Intelligence and Black-Box Medical Decisions: Accuracy versus Explainability. *Hastings Center Report*, *49*(1), 15–21. Retrieved 2022-11-10, from https://onlinelibrary.wiley.com/doi/abs/10.1002/hast.973 (_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/hast.973) doi: 10.1002/hast.973

Lorenz-Spreen, P., Oswald, L., Lewandowsky, S., & Hertwig, R. (2023). A systematic review of worldwide causal and correlational evidence on digital media and democracy. *Nature human behaviour*, *7*(1), 74–101.

Loveridge, R., Sallu, S. M., Pesha, I. J., & R Marshall, A. (2020, December). Measuring human wellbeing: A protocol for selecting local indicators. *Environmental Science & Policy*, *114*, 461–469. Retrieved 2023-04-03, from https://www.sciencedirect.com/science/article/pii/S146290112030112X doi: 10.1016/j.envsci.2020.09.002

Lu, F., Dumitrache, A., & Graus, D. (2020). Beyond optimizing for clicks: Incorporating editorial values in news recommendation. In *Proceedings of the 28th acm conference on user modeling, adaptation and personalization* (pp. 145–153).

Lukoff, K., Lyngs, U., Shirokova, K., Rao, R., Tian, L., Zade, H., … Hiniker, A. (2023). Switchtube: A proof-of-concept system introducing "adaptable commitment interfaces" as a tool for digital wellbeing. In *Proceedings of the 2023 chi conference on human factors in computing systems* (pp. 1–22).

Lynch, S. (2023, April). *AI Benchmarks Hit Saturation.* Stanford Institute for Human-Centered AI. Retrieved from https://hai.stanford.edu/news/ai-benchmarks-hit-saturation (Retrieved 2024-01-15)

Lyngs, U., Lukoff, K., Slovak, P., Seymour, W., Webb, H., Jirotka, M., … Shadbolt, N. (2020). 'i just want to hack myself to not get distracted' evaluating design interventions for self-control on facebook.

In *Proceedings of the 2020 chi conference on human factors in computing systems* (pp. 1–15).

Lyubomirsky, S. (2008). *The how of happiness: A scientific approach to getting the life you want.* penguin.

Lyubomirsky, S., & Abbe, A. (2003). Positive psychology's legs. *Psychological Inquiry*, *14*(2), 132–136.

Lyubomirsky, S., & Layous, K. (2013). How Do Simple Positive Activities Increase.
doi: 10.1177/0963721412469809

Lyubomirsky, S., & Lepper, H. S. (1999). A measure of subjective happiness: Preliminary reliability and construct validation. *Social indicators research*, *46*, 137–155.

Löhr, K., Weinhardt, M., & Sieber, S. (2020, January). The "World Café" as a Participatory Method for Collecting Qualitative Data. *International Journal of Qualitative Methods*, *19*, 1609406920916976. Retrieved 2022-06-20, from https://doi.org/10.1177/1609406920916976 (Publisher: SAGE Publications Inc) doi: 10.1177/1609406920916976

Mackenzie, S. B., Podsakoff, P. M., & Podsakoff, N. P. (2011). Linked references are available on JSTOR for this article : Construct Measurement and Validation Procedures in MIS and Behavioral Research : Integrating New and Existing Techniques1. *MIS Quarterly*, *35*(2), 293–334.

Madianou, M. (2021). Nonhuman humanitarianism: when'ai for good'can be harmful. *Information, Communication & Society*, *24*(6), 850–868.

Martelaro, N., & Ju, W. (2018, oct). Cybernetics and the design of the user experience of ai systems. *Interactions*, *25*(6), 38–41. doi: 10.1145/3274570

McGregor, J. A., Camfield, L., & Coulthard, S. (2015). Competing interpretations: human wellbeing and the use of quantitative and qualitative methods. *Mixed methods research in poverty and vulnerability: sharing ideas and learning lessons*, 231–260.

McShane, T. O., Hirsch, P. D., Trung, T. C., Songorwa, A. N., Kinzig, A., Monteferri, B., … O'Connor, S. (2011, March). Hard choices: Making trade-offs between biodiversity conservation and human well-being. *Biological Conservation*, *144*(3), 966–972. Retrieved 2024-01-03, from https://www.sciencedirect.com/science/article/pii/S0006320710001849 doi: 10.1016/j.biocon.2010.04.038

Mead, J., Fisher, Z., & Kemp, A. H. (2021). Moving Beyond Disciplinary Silos Towards a Transdisciplinary Model of Wellbeing: An Invited Review. *Frontiers in Psychology*, *12*. Retrieved 2023-12-18, from

https://www.frontiersin.org/articles/10.3389/fpsyg.2021.642093
Mead, M. (1968). *Cybernetics of cybernetics*. éditeur non identifié.
Mildner, T., Savino, G.-L., Doyle, P. R., Cowan, B. R., & Malaka, R. (2023). About engaging and governing strategies: A thematic analysis of dark patterns in social networking services. In *Proceedings of the 2023 chi conference on human factors in computing systems* (pp. 1–15).
Mindell, D. A. (2000). Cybernetics knowledge domains in engineering systems. *Research paper, Massachusetts Institute of Technology*.
Mittelstadt, B. (2019). Principles alone cannot guarantee ethical ai. *Nature machine intelligence*, *1*(11), 501–507.
Mizobuchi, H. (2014). Measuring World Better Life Frontier: A Composite Indicator for OECD Better Life Index. *Social Indicators Research*, *118*(3), 987–1007. Retrieved 2022-06-20, from https://www.jstor.org/stable/24721048 (Publisher: Springer)
Mohamed, S., Png, M.-T., & Isaac, W. (2020, December). Decolonial AI: Decolonial Theory as Sociotechnical Foresight in Artificial Intelligence. *Philosophy & Technology*, *33*(4), 659–684. Retrieved 2023-10-12, from https://link.springer.com/10.1007/s13347-020-00405-8 doi: 10.1007/s13347-020-00405-8
Möller, J., Trilling, D., Helberger, N., & van Es, B. (2020). Do not blame it on the algorithm: an empirical assessment of multiple recommender systems and their impact on content diversity. In *Digital media, political polarization and challenges to democracy* (pp. 45–63). Routledge.
Mollick, E. (2023, February). *Meet the teacher requiring students to use ChatGPT*. Retrieved 2024-01-17, from https://www.thirteen.org/metrofocus/2023/02/artificial-intelligence-school-scwlvl/
Monge Roffarello, A., & De Russis, L. (2019). The race towards digital wellbeing: Issues and opportunities. In *Proceedings of the 2019 chi conference on human factors in computing systems* (pp. 1–14).
Monge Roffarello, A., Lukoff, K., & De Russis, L. (2023). Defining and identifying attention capture deceptive designs in digital interfaces. In *Proceedings of the 2023 chi conference on human factors in computing systems* (pp. 1–19).
Moons, P., Budts, W., & De Geest, S. (2006, September). Critique on the conceptualisation of quality of life: A review and evaluation of different conceptual approaches. *International Journal of Nursing Studies*, *43*(7), 891–901. Retrieved 2022-12-12, from https://www.sciencedirect.com/science/article/pii/S0020748906001088 doi: 10.1016/j.ijnurstu.2006.03.015
Morley, J., Elhalal, A., Garcia, F., Kinsey, L., Mökander, J., & Floridi, L.

(2021, June). Ethics as a Service: A Pragmatic Operationalisation of AI Ethics. *Minds and Machines*, *31*(2), 239–256. Retrieved 2023-06-05, from https://doi.org/10.1007/s11023-021-09563-w doi: 10.1007/s11023-021-09563-w

Morley, J., Floridi, L., Kinsey, L., & Elhalal, A. (2020). From What to How: An Initial Review of Publicly Available AI Ethics Tools, Methods and Research to Translate Principles into Practices. *Science and Engineering Ethics*, *26*(4), 2141–2168. Retrieved from https://doi.org/10.1007/s11948-019-00165-5 (Publisher: Springer Netherlands) doi: 10.1007/s11948-019-00165-5

Morley, J., Kinsey, L., Elhalal, A., Garcia, F., Ziosi, M., & Floridi, L. (2021). Operationalising AI ethics: barriers, enablers and next steps. *AI & SOCIETY*, *38*(1), 411–423. Retrieved 2023-06-03, from https://doi.org/10.1007/s00146-021-01308-8 doi: 10.1007/s00146-021-01308-8

Mosseri, A. (2018, January 11). *Bringing people closer together.* Facebook Newsroom. (URL: https://newsroom.fb.com/news/2018/01/news-feed-fyi-bringing-people-closer-together/ (Retrieved 2024-01-04))

Musikanski, L., Rakova, B., Bradbury, J., Phillips, R., & Manson, M. (2020). Artificial Intelligence and Community Well-being: A Proposal for an Emerging Area of Research. *International Journal of Community Well-Being*, *3*(1), 39–55. (Publisher: International Journal of Community Well-Being) doi: 10.1007/s42413-019-00054-6

Müller, K., & Schwarz, C. (2018). Fanning the flames of hate: Social media and hate crime. *Journal of the European Economic Association*, *19*(2), 2131–2167.

Ng, A. Y., & Russell, S. J. (2000). Algorithms for inverse reinforcement learning. In *Proceedings of the seventeenth international conference on machine learning* (p. 663–670). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Ng, D. T. K., Leung, J. K. L., Chu, S. K. W., & Qiao, M. S. (2021). Conceptualizing ai literacy: An exploratory review. *Computers and Education: Artificial Intelligence*, *2*, 100041.

Nielsen, J. (1994). *Usability engineering.* Morgan Kaufmann.

Noë, A. (2004). *Action in perception.* MIT press.

Norem, J. K., & Chang, E. C. (2002). The positive psychology of negative thinking. *Journal of clinical psychology*, *58*(9), 993–1001.

Norman, D. A. (2005, July). Human-centered design considered harmful. *Interactions*, *12*(4), 14–19. Retrieved 2022-05-04, from https://dl.acm.org/doi/10.1145/1070960.1070976 doi: 10.1145/1070960.1070976

Norman, D. A. (2013). *The design of everyday things.* MIT Press,

Cambridge.

Norman, D. A., & Stappers, P. J. (2015). DesignX: Complex Sociotechnical Systems. *She Ji*, *1*(2), 83–106. doi: 10.1016/j.sheji.2016.01.002

O'Mahony, T. (2022). Toward Sustainable Wellbeing: Advances in Contemporary Concepts. *Frontiers in Sustainability*, *3*. Retrieved 2024-01-04, from https://www.frontiersin.org/articles/10.3389/frsus.2022.807984

OpenAI. (2022, Jan). *Aligning language models to follow instructions.* OpenAI. Retrieved from https://openai.com/research/instruction-following

Organization, W. H. (n.d.). *World mental health report: Transforming mental health for all* (Tech. Rep.). Retrieved 2023-04-04, from https://www.who.int/publications-detail-redirect/9789240049338

Orlowski, J. (2020). *The Social Dilemma.* Retrieved 2022-05-04, from https://www.netflix.com/watch/81254224?trackId=255824129&tctx=0%2C0%2CNAPA%40%40%7Cee90d9c9-8034-4c5f-b248-ce517c583263-14092566_titles%2F1%2F%2Fthe%20social%20dilemma%2F0%2F0%2CNAPA%40%40%7Cee90d9c9-8034-4c5f-b248-ce517c583263-14092566_titles%2F1%2F%2Fthe%20social%20dilemma%2F0%2F0%2Cunknown%2C%2Cee90d9c9-8034-4c5f-b248-ce517c583263-14092566%7C1%2CtitlesResults

Ozmen Garibay, O., Winslow, B., Andolina, S., Antona, M., Bodenschatz, A., Coursaris, C., … Xu, W. (2023, February). Six Human-Centered Artificial Intelligence Grand Challenges. *International Journal of Human–Computer Interaction*, *39*(3), 391–437. Retrieved 2023-01-10, from https://doi.org/10.1080/10447318.2022.2153320 (Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/10447318.2022.2153320) doi: 10.1080/10447318.2022.2153320

Pan, A., Bhatia, K., & Steinhardt, J. (2022). The effects of reward misspecification: Mapping and mitigating misaligned models. In *International conference on learning representations.* Retrieved from https://openreview.net/forum?id=JYtwGwIL7ye

Pangaro, P. (2017). *Cybernetics as Phoenix: Why Ashes, What New Life* (Vol. 1). Retrieved from http://link.springer.com/10.1007/978-3-642-01310-2_2 (Publication Title: Conversations. Cybernetics: State of the Art)

Pangaro, P. (2021, August). *#NewMacy 2021: Responding to Pandemics of "Today's AI".* Retrieved 2022-05-04, from https://pangaro.com/designconversation/2021/08/newmacy-in-2021-pandemics-ai/

Papanek, V., & Fuller, R. B. (1972). Design for the real world.

Parisi, L., & Comunello, F. (2020). Dating in the time of "relational filter bubbles": Exploring imaginaries, perceptions and tactics of italian dating app users. *The Communication Review*, *23*(1), 66–89.

Pavot, W. (2014). Temporal Satisfaction with Life Scale (TSWLS). In A. C. Michalos (Ed.), *Encyclopedia of Quality of Life and Well-Being Research* (pp. 6609–6611). Dordrecht: Springer Netherlands. Retrieved 2023-03-03, from https://doi.org/10.1007/978-94-007-0753-5_2993 doi: 10.1007/978-94-007-0753-5_2993

Pavot, W., & Diener, E. (2008). The satisfaction with life scale and the emerging construct of life satisfaction. *The journal of positive psychology*, *3*(2), 137–152.

Peters, D. (2023). Wellbeing supportive design–research-based guidelines for supporting psychological wellbeing in user experience. *International Journal of Human–Computer Interaction*, *39*(14), 2965–2977.

Peters, D., Calvo, R. A., & Ryan, R. M. (2018). Designing for motivation, engagement and wellbeing in digital experience. *Frontiers in Psychology*, *9*(MAY), 1–15. doi: 10.3389/fpsyg.2018.00797

Peters, D., Sadek, M., & Ahmadpour, N. (2023). Collaborative workshops at scale: A method for non-facilitated virtual collaborative design workshops. *International Journal of Human–Computer Interaction*, 1–18.

Peters, D., Vold, K., Robinson, D., & Calvo, R. A. (2020). Responsible ai—two frameworks for ethical design practice. *IEEE Transactions on Technology and Society*, *1*(1), 34–47.

Phillips, R., & Wong, C. (Eds.). (2017). *Handbook of Community Well-Being Research*. Dordrecht: Springer Netherlands. Retrieved 2022-07-14, from http://link.springer.com/10.1007/978-94-024-0878-2 doi: 10.1007/978-94-024-0878-2

Piet, N., & MOBGEN. (2019). *Ai meets design toolkit.* https://aixdesign.gumroad.com/l/toolkit. (Toolkit for designing human-centered AI applications. Includes crash course in AI/ML, prompts, exercises, worksheets, and guides for feasibility, viability, and desirability assessments.)

Piliavin, J. A., & Siegl, E. (2007). Health benefits of volunteering in the Wisconsin longitudinal study. *Journal of Health and Social Behavior*, *48*(4), 450–464. doi: 10.1177/002214650704800408

Piorkowski, D., Park, S., Wang, A. Y., Wang, D., Muller, M., & Portnoy, F. (2021). How ai developers overcome communication challenges in a multidisciplinary team: A case study. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW1), 1–25.

Prabhakaran, V., Mitchell, M., Gebru, T., & Gabriel, I. (2022, October). *A Human Rights-Based Approach to Responsible AI.* arXiv. Retrieved 2023-12-20, from http://arxiv.org/abs/2210.02667 (arXiv:2210.02667 [cs])

Pressman, A. (2019). *Design Thinking: A Guide to Creative Problem Solving for Everyone.* Routledge. (Google-Books-ID: mygCugEACAAJ)

Pretorius, T.-l. (2021, June). Depression among health care students in the time of COVID-19: the mediating role of resilience in the hopelessness–depression relationship. *South African Journal of Psychology*, *51*(2), 269–278. Retrieved 2022-06-01, from https://doi.org/10.1177/0081246321994452 (Publisher: SAGE Publications) doi: 10.1177/0081246321994452

Qasem, F. (2023, January). ChatGPT in scientific and academic research: future fears and reassurances. *Library Hi Tech News*, *40*(3), 30–32. Retrieved 2023-12-20, from https://doi.org/10.1108/LHTN-03-2023-0043 (Publisher: Emerald Publishing Limited) doi: 10.1108/LHTN-03-2023-0043

Radler, B. T., & Ryff, C. D. (2010). Who Participates? Accounting for Longitudinal Retention in the MIDUS National Study of Health and Well-Being. *Journal of Aging and Health*, *22*(3), 307–331. doi: 10.1177/0898264309358617

Raibley, J. R. (2012). Happiness is not well-being. *Journal of Happiness Studies*, *13*, 1105–1129.

Rambur, B., Vallett, C., Cohen, J. A., & Tarule, J. M. (2013, November). Metric-driven harm: An exploration of unintended consequences of performance measurement. *Applied Nursing Research*, *26*(4), 269–272. Retrieved 2022-06-20, from https://www.sciencedirect.com/science/article/pii/S0897189713000815 doi: 10.1016/j.apnr.2013.09.001

Rana, N. P., Chatterjee, S., Dwivedi, Y. K., & Akter, S. (2022). Understanding dark side of artificial intelligence (ai) integrated business analytics: assessing firm's operational inefficiency and competitiveness. *European Journal of Information Systems*, *31*(3), 364–387. doi: 10.1080/0960085X.2021.1955628

Renger, R. F., Midyett, S. J., Soto Mas, F. G., Erin, T. D., McDermott, H. M., Papenfuss, R. L., … Hewitt, M. J. (2000). Optimal living profile: An inventory to assess health and wellness. *American journal of health behavior*, *24*(6), 403–412.

Renshaw, T. L., & Bolognino, S. J. (2014). The College Student Subjective Wellbeing Questionnaire : A Brief , Multidimensional Measure of Undergraduate ' s Covitality.

doi: 10.1007/s10902-014-9606-4

Renshaw, T. L., Long, A. C. J., & Cook, C. R. (2014). Assessing Adolescents ' Positive Psychological Functioning at School : Development and Validation of the Student Subjective Wellbeing Questionnaire. , *29*(3).

Rimé, B., Bouchat, P., Paquette, C., & Mesquita, B. (2019). Intrapersonal, interpersonal, and social outcomes of the social sharing of emotion. *Current Opinion in Psychology*, *31*, 127–134.

Ruggeri, K., Garcia-Garzon, E., Maguire, Á., Matz, S., & Huppert, F. A. (2020). Well-being is more than happiness and life satisfaction: a multidimensional analysis of 21 countries. *Health and quality of life outcomes*, *18*(1), 1–16.

Russell, S. J., & Norvig, P. (2022). *Artificial intelligence: a modern approach* (Fourth edition, global edition ed.). Harlow: Pearson.

Ryan, R. M., & Deci, E. L. (2001). On happiness and human potentials: A review of research on hedonic and eudaimonic well-being. *Annual Review of Psychology*, *52*, 141–166. doi: 10.1146/annurev.psych.52.1.141

Ryff, C. D. (1989). Happiness is everything, or is it? Explorations on the meaning of psychological well-being. *Journal of Personality and Social Psychology*, *57*(6), 1069–1081. doi: 10.1037/0022-3514.57.6.1069

Ryff, C. D., & Keyes, C. L. M. (1995). The Structure of Psychological Well-Being Revisited. *Journal of Personality and Social Psychology*, *69*(4), 719–727. doi: 10.1037/0022-3514.69.4.719

Ryff, C. D., & Singer, B. (2003). Ironies of the human condition: Well-being and health on the way to mortality.

Ryff, C. D., & Singer, B. H. (2006). Best news yet on the six-factor model of well-being. *Social Science Research*, *35*(4), 1103–1119. doi: 10.1016/j.ssresearch.2006.01.002

Sachdeva, S., Singh, P., & Medin, D. (2011). Culture and the quest for universal principles in moral reasoning. *International Journal of Psychology*, *46*(3), 161–176.

Sadek, M., Calvo, R. A., & Mougenot, C. (2023a). Co-designing conversational agents: A comprehensive review and recommendations for best practices. *Design Studies*, *89*, 101230.

Sadek, M., Calvo, R. A., & Mougenot, C. (2023b). Designing value-sensitive ai: a critical review and recommendations for socio-technical design processes. *AI and Ethics*. Retrieved from https://doi.org/10.1007/s43681-023-00373-7 doi: 10.1007/s43681-023-00373-7

Sadek, M., Calvo, R. A., & Mougenot, C. (2023c, October). *The Value-Sensitive Conversational Agent Co-Design Framework.*

arXiv. Retrieved 2023-10-19, from http://arxiv.org/abs/2310.11848 (arXiv:2310.11848 [cs])

Sanches, P., Janson, A., Karpashevich, P., Nadal, C., Qu, C., Daudén Roquet, C., … Sas, C. (2019, May). HCI and Affective Health: Taking stock of a decade of studies and charting future research directions. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (pp. 1–17). New York, NY, USA: Association for Computing Machinery. Retrieved 2022-06-01, from https://doi.org/10.1145/3290605.3300475 doi: 10.1145/3290605.3300475

Sanders, E. B.-N., & Stappers, P. J. (2008). Co-creation and the new landscapes of design. *Co-design*, *4*(1), 5–18.

Sanderson, C., Douglas, D., Lu, Q., Schleiger, E., Whittle, J., Lacey, J., … Hansen, D. (2023). Ai ethics principles in practice: Perspectives of designers and developers. *IEEE Transactions on Technology and Society*.

Sartori, L., & Theodorou, A. (2022). A sociotechnical perspective for the future of ai: narratives, inequalities, and human control. *Ethics and Information Technology*, *24*(1), 4.

Sas, C., Whittaker, S., Dow, S., Forlizzi, J., & Zimmerman, J. (2014). Generating implications for design through design research. *Conference on Human Factors in Computing Systems - Proceedings*, 1971–1980. doi: 10.1145/2556288.2557357

Sato, K. (1991). From ai to cybernetics. *AI & society*, *5*, 155–161. Retrieved from https://doi.org/10.1007/BF01891721

Schiff, D., Ayesh, A., Musikanski, L., & Havens, J. C. (2020). IEEE 7010: A New Standard for Assessing the Well-being Implications of Artificial Intelligence. *Conference Proceedings - IEEE International Conference on Systems, Man and Cybernetics*, *2020-Octob*(March), 2746–2753. doi: 10.1109/SMC42975.2020.9283454

Schiff, D., Biddle, J., Borenstein, J., & Laas, K. (2020). What's next for ai ethics, policy, and governance? a global overview. In *Proceedings of the aaai/acm conference on ai, ethics, and society* (pp. 153–158).

Schiff, D., Borenstein, J., Biddle, J., & Laas, K. (2021). Ai ethics in the public, private, and ngo sectors: A review of a global document collection. *IEEE Transactions on Technology and Society*, *2*(1), 31–42.

Schiff, D., Murahwi, Z., Musikanski, L., & Havens, J. (2019). A new paradigm for autonomous and intelligent systems development: Why well-being measurement matters. In *Workshop on designing digital wellbeing, chi 2019.*

Schiff, D., Rakova, B., Ayesh, A., Fanti, A., & Lennon, M. (2020). Principles to practices for responsible ai: closing the gap. *arXiv preprint arXiv:2006.04707*.

Schiff, D., Rakova, B., Ayesh, A., Fanti, A., & Lennon, M. (2021). Explaining the principles to practices gap in ai. *IEEE Technology and Society Magazine*, *40*(2), 81–94.

Schwab, K. (2017). *The Fourth Industrial Revolution*. Crown. (Google-Books-ID: ST_FDAAAQBAJ)

Schwartz, S. H., et al. (2012). An overview of the schwartz theory of basic values. *Online readings in Psychology and Culture*, *2*(1), 2307–0919.

Scott, A. J., Webb, T. L., Martyn-St James, M., Rowse, G., & Weich, S. (2021, December). Improving sleep quality leads to better mental health: A meta-analysis of randomised controlled trials. *Sleep Medicine Reviews*, *60*, 101556. doi: 10.1016/j.smrv.2021.101556

Scott, B. (2001). Gordon pask's conversation theory: A domain independent constructivist model of human knowing. *Foundations of Science*, *6*, 343–360.

Scott, B. (2004). Second-order cybernetics: an historical introduction. *Kybernetes*, *33*(9/10), 1365–1378.

Searle, B. A., Pykett, J., & Alfaro-Simmonds, M. J. (2021, June). Introduction to wellbeing research. In *A Modern Guide to Well-being Research* (pp. 1–21). Edward Elgar Publishing. Retrieved 2023-12-23, from https://china.elgaronline.com/edcollchap/edcoll/9781789900156/9781789900156.00008.xml (Section: A Modern Guide to Wellbeing Research)

Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. In *Proceedings of the conference on fairness, accountability, and transparency* (p. 59–68). New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi.org/10.1145/3287560.3287598 doi: 10.1145/3287560.3287598

Seligman, M. E. (1999). The president's address. *American psychologist*, *54*(8), 559–562. doi: 10.1037/0003-066X.54.8.537

Seligman, M. E. (2010). Flourish: Positive Psychology and Positive Interventions. *The Tanner Lectures on Human Values*, 231–242.

Seligman, M. E. (2011). *Flourish: a visionary new understanding of happiness and well-being*. Policy. (Pages: 349)

Seligman, M. E. (2019). Positive psychology: A personal history. *Annual review of clinical psychology*, *15*, 1–23.

Seligman, M. E., & Csikszentmihalyi, M. (2000). *Positive psychology: An*

introduction. (Vol. 55) (No. 1). American Psychological Association.

Shah, A. M., Coordinator, P., Williams, C., Investigator, C.-p., Delgado, J., Design, S., … Investigator, P. (2003). A Participatory Approach to Designing a Community Health Survey A Report on the Survey Development Process. (043026).

Shahriari, K., & Shahriari, M. (2017). IEEE standard review - Ethically aligned design: A vision for prioritizing human wellbeing with artificial intelligence and autonomous systems. *IHTC 2017 - IEEE Canada International Humanitarian Technology Conference 2017*, 197–201. doi: 10.1109/IHTC.2017.8058187

Shapiro, M. A., Barriga, C. A., & Beren, J. (2010). Causal attribution and perceived realism of stories. *Media Psychology*, *13*(3), 273–300.

Shen, T., Jin, R., Huang, Y., Liu, C., Dong, W., Guo, Z., … Xiong, D. (2023). Large language model alignment: A survey. *arXiv preprint arXiv:2309.15025*.

Shneiderman, B. (2022). *Human-Centered AI.* Oxford University Press. (Google-Books-ID: YS9VEAAAQBAJ)

Sirgy, M. J. (2018, June). The Psychology of Material Well-Being. *Applied Research in Quality of Life*, *13*(2), 273–301. Retrieved 2022-08-03, from https://doi.org/10.1007/s11482-017-9590-z doi: 10.1007/s11482-017-9590-z

Sone, T., Nakaya, N., Ohmori, K., Shimazu, T., Higashiguchi, M., Kakizaki, M., … Tsuji, I. (2008). Sense of life worth living (ikigai) and mortality in japan: Ohsaki study. *Psychosomatic medicine*, *70*(6), 709–715.

Spektor, F., Fox, S. E., Awumey, E., Riordan, C. A., Rho, H. J., Kulkarni, C., … Forlizzi, J. (2023). Designing for wellbeing: Worker-generated ideas on adapting algorithmic management in the hospitality industry. In *Proceedings of the 2023 acm designing interactive systems conference* (pp. 623–637).

Spotify. (2023, June 20). *Spotify's desktop experience gets a brand-new look with redesigned 'your library' and 'now playing' views.* Spotify Newsroom. Retrieved from https://newsroom.spotify.com/2023-06-20/spotifys-desktop-experience-gets-a-brand-new-look-with-redesigned-your-library-and-now-playing-views/ (Retrieved 2023-01-12)

Springer, K. W., & Hauser, R. M. (2006). An assessment of the construct validity of Ry V ' s Scales of Psychological Well-Being : Method , mode , and measurement e V ects. , *35*, 1080–1102. doi: 10.1016/j.ssresearch.2005.07.004

Stappers, P. J., & Giaccardi, E. (2017). Research through design. In *The encyclopedia of human-computer interaction* (pp. 1–94). The

Interaction Design Foundation.

Stephen, D. (2021, July). *Pace layers in experience design — Stabilise innovation by understanding people's needs.* Retrieved 2023-02-07, from https://duncanstephen.net/pace-layers-in-experience-design-stabilise-innovation-by-understanding-peoples-needs/ (Section: Design)

Sternberg, R. J. (2003). A broad view of intelligence: the theory of successful intelligence. *Consulting Psychology Journal: Practice and Research*, *55*(3), 139.

Steur, A. J., & Seiter, M. (2021, May). Properties of feedback mechanisms on digital platforms: an exploratory study. *Journal of Business Economics*, *91*(4), 479–526. Retrieved 2023-12-13, from https://doi.org/10.1007/s11573-020-01009-6 doi: 10.1007/s11573-020-01009-6

Stocké, V., & Langfeldt, B. (2004). Effects of survey experience on respondents' attitudes towards surveys. *Bulletin de Méthodologie Sociologique*, *81*(1), 5–32. doi: 10.1177/075910630408100103

Stray, J. (2020). Aligning AI Optimization to Community Well-Being. *International Journal of Community Well-Being*, *3*(4), 443–463. (Publisher: International Journal of Community Well-Being) doi: 10.1007/s42413-020-00086-3

Stray, J., & Hadfield, G. K. (n.d.). Platforms Can Optimize for Metrics Beyond Engagement. *Wired*. Retrieved 2023-03-03, from https://www.wired.com/story/platforms-engagement-research-meta/ (Section: tags)

Stray, J., Halevy, A., Assar, P., Hadfield-Menell, D., Boutilier, C., Ashar, A., … others (2022). Building human values into recommender systems: An interdisciplinary synthesis. *ACM Transactions on Recommender Systems*.

Stray, J., Halevy, A., Assar, P., Hadfield-Menell, D., Boutilier, C., Ashar, A., … Vasan, N. (2023, November). Building Human Values into Recommender Systems: An Interdisciplinary Synthesis. *ACM Transactions on Recommender Systems*, 3632297. Retrieved 2023-12-05, from https://dl.acm.org/doi/10.1145/3632297 doi: 10.1145/3632297

Stsiampkouskaya, K., Joinson, A., Piwek, L., & Stevens, L. (2021). Imagined audiences, emotions, and feedback expectations in social media photo sharing. *Social Media+ Society*, *7*(3), 20563051211035692.

Subramonyam, H., Seifert, C., & Adar, E. (2021). Towards a process model for co-creating ai experiences. In *Proceedings of the 2021 acm designing interactive systems conference* (pp. 1529–1543).

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement Learning, second edition:*

An Introduction. MIT Press. (Google-Books-ID: sWV0DwAAQBAJ)

Sweeting, B. (2016). Design research as a variety of second-order Cybernetic practice. *Constructivist Foundations*, *11*(3), 572–579. doi: 10.1142/9789813226265_0035

Tabari, P. (2022, September). The Role of Artificial Intelligence in Human-Computer Interaction: Using a Smart Topic Extraction System. In *2022 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)* (pp. 1–3). (ISSN: 1943-6106) doi: 10.1109/VL/HCC53370.2022.9833114

Tan, C. (2020, December). The impact of COVID-19 on student motivation, community of inquiry and learning performance. *Asian Education and Development Studies*, *10*(2), 308–321. Retrieved 2022-06-01, from https://www.emerald.com/insight/content/doi/10.1108/AEDS-05-2020-0084/full/html doi: 10.1108/AEDS-05-2020-0084

Technicalities, & Stag. (2023, November). *Shallow review of live agendas in alignment & safety — AI Alignment Forum.* Retrieved 2023-12-22, from https://www.alignmentforum.org/posts/zaaGsFBeDTpCsYHef/shallow-review-of-live-agendas-in-alignment-and-safety

Tegmark, M. (2023, April). *The 'Don't Look Up' Thinking That Could Doom Us With AI.* Retrieved 2023-10-05, from https://time.com/6273743/thinking-that-could-doom-us-with-ai/

Tennant, R., Hiller, L., Fishwick, R., Platt, S., Joseph, S., Weich, S., … Stewart-Brown, S. (2007). The Warwick-Dinburgh mental well-being scale (WEMWBS): Development and UK validation. *Health and Quality of Life Outcomes*, *5*, 1–13. doi: 10.1186/1477-7525-5-63

Thomas, R. L., & Uminsky, D. (2020). Reliance on Metrics is a Fundamental Challenge for AI. *Ethics of Data Science Conference*. Retrieved from https://arxiv.org/abs/2002.08512

Thorburn, L. (2022, August). *How to Measure the Causal Effects of Recommenders.* Retrieved 2022-08-24, from https://medium.com/understanding-recommenders/how-to-measure-the-causal-effects-of-recommenders-5e89b7363d57

Tomašev, N., Cornebise, J., Hutter, F., Mohamed, S., Picciariello, A., Connelly, B., … others (2020). Ai for social good: unlocking the opportunity for positive impact. *Nature Communications*, *11*(1), 2468.

Tommasel, A., Rodriguez, J. M., & Godoy, D. (2022). Haven't i just listened to this?: Exploring diversity in music recommendations. In *Adjunct proceedings of the 30th acm conference on user modeling, adaptation and personalization* (pp. 35–40).

Topp, C. W., Østergaard, S. D., Søndergaard, S., & Bech, P. (2015).

The WHO-5 well-being index: A systematic review of the literature. *Psychotherapy and Psychosomatics*, *84*(3), 167–176. doi: 10.1159/000376585

Tromp, N., & Hekkert, P. (2016, March). Assessing methods for effect-driven design: Evaluation of a social design method. *Design Studies*, *43*, 24–47. Retrieved 2023-01-24, from https://linkinghub.elsevier.com/retrieve/pii/S0142694X15000976 doi: 10.1016/j.destud.2015.12.002

Tromp, N., & Hekkert, P. (2019). *Designing for Society: Products and Services for a Better World*. Bloomsbury Academic.

Trunfio, M., & Rossi, S. (2021, September). Conceptualising and measuring social media engagement: A systematic literature review. *Italian Journal of Marketing*, *2021*(3), 267–292. Retrieved 2023-03-03, from https://doi.org/10.1007/s43039-021-00035-8 doi: 10.1007/s43039-021-00035-8

Turchin, A., & Denkenberger, D. (2020). Classification of global catastrophic risks connected with artificial intelligence. *Ai & Society*, *35*(1), 147–163.

Uchida, Y., & Kitayama, S. (2009). Happiness and unhappiness in east and west: themes and variations. *Emotion*, *9*(4), 441.

Ulanoff, L. (2023, November). *OpenAI's reported 'superintelligence' breakthrough is so big it nearly destroyed the company, and ChatGPT.* Retrieved 2023-12-12, from https://www.techradar.com/computing/artificial-intelligence/openais-reported-superintelligence-breakthrough-is-so-big-it-nearly-destroyed-the-company-and-chatgpt

Umbrello, S., & De Bellis, A. F. (2018). A Value-Sensitive Design Approach to Intelligent Agents. *Artificial Intelligence Safety and Security*(January), 395–410. doi: 10.13140/RG.2.2.17162.77762

Umbrello, S., & van de Poel, I. (2021, August). Mapping value sensitive design onto AI for social good principles. *AI and Ethics*, *1*(3), 283–296. Retrieved 2022-05-11, from https://link.springer.com/10.1007/s43681-021-00038-3 doi: 10.1007/s43681-021-00038-3

Ura, K., Alkire, S., & Zangmo, T. (2012). GNH and GNH Index. *The Centre for Bhutan Studies*(May), 1–60.

Valencia, A. L., & Froese, T. (2020). What binds us? inter-brain neural synchronization and its implications for theories of human consciousness. *Neuroscience of consciousness*, *2020*(1), niaa010.

van der Maden, W., Lomas, D., & Hekkert, P. (2022). Design for wellbeing during covid-19: A cybernetic perspective on data feedback loops in complex sociotechnical systems. In D. Lockton, S. Lenzi, P. Hekkert, A. Oak, J. Sádaba, & P. Lloyd (Eds.), *Drs2022: Bilbao*. Bilbao, Spain.

doi: 10.21606/drs.2022.771

van der Maden, W., Lomas, D., Sadek, M., & Hekkert, P. (2024). *Positive AI: Key challenges in designing artificial intelligence for wellbeing.*

van Allen, P. (2018, oct). Prototyping ways of prototyping ai. *Interactions*, *25*(6), 46–51. Retrieved from https://doi.org/10.1145/3274566 doi: 10.1145/3274566

Van Berkel, N., Ferreira, D., & Kostakos, V. (2017). The experience sampling method on mobile devices. *ACM Computing Surveys (CSUR)*, *50*(6), 1–40.

Vanden Abeele, M. M. P. (2021, November). Digital Wellbeing as a Dynamic Construct. *Communication Theory*, *31*(4), 932–955. Retrieved 2022-09-07, from https://doi.org/10.1093/ct/qtaa024 doi: 10.1093/ct/qtaa024

van de Poel, I. (2020, September). Embedding Values in Artificial Intelligence (AI) Systems. *Minds and Machines*, *30*(3), 385–409. Retrieved 2022-12-02, from https://doi.org/10.1007/s11023-020-09537-4 doi: 10.1007/s11023-020-09537-4

van der Maden, W., Lomas, D., & Hekkert, P. (2023). A framework for designing AI systems that support community wellbeing. *Frontiers in Psychology*, *13*. Retrieved 2023-01-15, from https://www.frontiersin.org/articles/10.3389/fpsyg.2022.1011883

van Dijk, J., & van der Lugt, R. (2013). Scaffolds for design communication: Research through design of shared understanding in design meetings. *Artificial Intelligence for Engineering Design, Analysis and Manufacturing*, *27*(2), 121–131. doi: 10.1017/S0890060413000024

van Zyl, L. E., Gaffaney, J., van der Vaart, L., Dik, B. J., & Donaldson, S. I. (2023). The critiques and criticisms of positive psychology: A systematic review. *The Journal of Positive Psychology*, 1–30.

Vardi, M. Y. (2012, jan). Artificial intelligence: Past and future. *Commun. ACM*, *55*(1), 5. Retrieved from https://doi.org/10.1145/2063176.2063177 doi: 10.1145/2063176.2063177

Varshney, K. R. (2023, September). *Decolonial AI Alignment: Vi\'{s}esadharma, Argument, and Artistic Expression.* arXiv. Retrieved 2023-10-12, from http://arxiv.org/abs/2309.05030 (arXiv:2309.05030 [cs, stat])

Vassilakopoulou, P. (2020). *Sociotechnical approach for accountability by design in AI systems.*

Veenhoven, R. (2014). Journal of Happiness Studies. *Encyclopedia of Quality of Life and Well-Being Research*, 3462–3464. doi: 10.1007/978-94-007-0753-5_4049

Vinuesa, R., Azizpour, H., Leite, I., Balaam, M., Dignum, V., Domisch, S., … Fuso Nerini, F. (2020, January). The role of artificial intelligence in achieving the Sustainable Development Goals. *Nature Communications*, *11*(1), 233. Retrieved 2024-01-04, from https://www.nature.com/articles/s41467-019-14108-y (Number: 1 Publisher: Nature Publishing Group) doi: 10.1038/s41467-019-14108-y

Virokannas, E., Liuski, S., & Kuronen, M. (2020, March). The contested concept of vulnerability – a literature review: Vulnerability-käsitteen kiistanalaiset merkitykset – systemaattinen kirjallisuuskatsaus. *European Journal of Social Work*, *23*(2), 327–339. Retrieved 2022-08-31, from https://www.tandfonline.com/doi/full/10.1080/13691457.2018.1508001 doi: 10.1080/13691457.2018.1508001

Visser, F. S., Stappers, P. J., Van der Lugt, R., & Sanders, E. B. (2005). Contextmapping: experiences from practice. *CoDesign*, *1*(2), 119–149.

Vogel, E. A., & Rose, J. P. (2016). Self-reflection and interpersonal connection: Making the most of self-presentation on social media. *Translational Issues in Psychological Science*, *2*(3), 294.

von Foerster, H. (2003). Cybernetics of Cybernetics. *Understanding Understanding*, 283–286. doi: 10.1007/0-387-21722-3_13

Voukelatou, V., Gabrielli, L., Miliou, I., Cresci, S., Sharma, R., Tesconi, M., & Pappalardo, L. (2021). Measuring objective and subjective well-being: dimensions and data sources. *International Journal of Data Science and Analytics*, *11*, 279–309.

Wach, K., Duong, C. D., Ejdys, J., Kazlauskaitė, R., Korzynski, P., Mazurek, G., … Ziemba, E. (2023). The dark side of generative artificial intelligence: A critical analysis of controversies and risks of chatgpt. *Entrepreneurial Business and Economics Review*, *11*(2), 7–24.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. R. (2018). Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

Wang, D., Churchill, E., Maes, P., Fan, X., Shneiderman, B., Shi, Y., & Wang, Q. (2020, April). From Human-Human Collaboration to Human-AI Collaboration: Designing AI Systems That Can Work Together with People. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems* (pp. 1–6). Honolulu HI USA: ACM. Retrieved 2023-02-20, from https://dl.acm.org/doi/10.1145/3334480.3381069 doi: 10.1145/3334480.3381069

Waterman, A. S. (1993). Two conceptions of happiness: Contrasts of personal expressiveness (eudaimonia) and hedonic enjoyment. *Journal of personality and social psychology*, *64*(4), 678.

Waterman, A. S., Schwartz, S. J., Zamboanga, B. L., Ravert, R. D., Williams, M. K., Bede Agocha, V., … Brent Donnellan, M. (2010). The questionnaire for eudaimonic well-being: Psychometric properties, demographic comparisons, and evidence of validity. *The journal of positive psychology*, *5*(1), 41–61.

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology*, *54*(6), 1063–1070. (Place: US Publisher: American Psychological Association) doi: 10.1037/0022-3514.54.6.1063

Weinstein, N. D., Marcus, S. E., & Moser, R. P. (2005). Smokers' unrealistic optimism about their risk. *Tobacco control*, *14*(1), 55–59.

Weizenbaum, J. (1966). Eliza—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, *9*(1), 36–45.

West, S., Whittaker, M., & Crawford, K. (2019). *Discriminating Systems: Gender, Race, and Power* (Tech. Rep.). AI Now Institute.

White, R. G., & Van Der Boor, C. (2020, September). Impact of the COVID-19 pandemic and initial period of lockdown on the mental health and well-being of adults in the UK. *BJPsych Open*, *6*(5), e90. Retrieved 2022-06-01, from https://www.cambridge.org/core/product/identifier/S2056472420000794/type/journal_article doi: 10.1192/bjo.2020.79

Wiblin, R. (n.d.). *Jan Leike on OpenAI's massive push to make superintelligence safe in 4 years or less.* Retrieved 2023-12-20, from https://80000hours.org/podcast/episodes/jan-leike-superalignment/

Wiener, N. (1961). *Cybernetics Or Control and Communication in the Animal and the Machine.* MIT Press.

Wiese, L., Pohlmeyer, A. E., & Hekkert, P. (2020, September). Design for Sustained Wellbeing through Positive Activities—A Multi-Stage Framework. *Multimodal Technologies and Interaction*, *4*(4), 71. Retrieved 2022-10-05, from https://www.mdpi.com/2414-4088/4/4/71 doi: 10.3390/mti4040071

Williams, G. M., Pendlebury, H., Thomas, K., & Smith, A. P. (2017). The Student Well-Being Process Questionnaire (Student WPQ). , 1748–1761. doi: 10.4236/psych.2017.811115

Wissing, M. P. (2022). Beyond the "third wave of positive psychology": Challenges and opportunities for future research. *Frontiers in Psychology*, *12*, 795067.

Wissing, M. P., Schutte, L., & Liversage, C. (2022). Embracing well-being

in diverse contexts: The third wave of positive psychology and african imprint. In *Embracing well-being in diverse african contexts: Research perspectives* (pp. 3–30). Springer.

Wong, C. S., & Law, K. S. (1999). Multidimensional constructs in structural equation analysis: An illustration using the job perception and job satisfaction constructs. *Journal of Management*, *25*(2), 143–160. doi: 10.1177/014920639902500202

Wong, P. T. (2011). Positive psychology 2.0: Towards a balanced interactive model of the good life. *Canadian Psychology/Psychologie canadienne*, *52*(2), 69–81. doi: 10.1037/a0022511

Wong, R., Madaio, M., & Merrill, N. (2022). Seeing like a toolkit: How toolkits envision the work of ai ethics. *Computing Research Repository (CoRR)*. Retrieved from https://arxiv.org/abs/2202.08792

World Health Organization, P. R. U. (1998). WHO (Five) Well-Being Index (1998 version). *WHO (Five) Well-Being Index (1998)*, 0–1. Retrieved from https://www.psykiatri-regionh.dk/who-5/Documents/WHO-5questionaire-English.pdf

Wragg, D. W. (1973). *A dictionary of aviation*. Osprey Publishing, Oxford.

Xiong, J., Lipsitz, O., Nasri, F., Lui, L. M., Gill, H., Phan, L., … McIntyre, R. S. (2020). Impact of COVID-19 pandemic on mental health in the general population: A systematic review. *Journal of Affective Disorders*, *277*(July), 55–64. Retrieved from https://doi.org/10.1016/j.jad.2020.08.001 (Publisher: Elsevier B.V.) doi: 10.1016/j.jad.2020.08.001

Yang, Q., Steinfeld, A., Rosé, C., & Zimmerman, J. (2020). Re-examining whether, why, and how human-ai interaction is uniquely difficult to design. In *Proceedings of the 2020 chi conference on human factors in computing systems* (pp. 1–13).

Yang, Q., Steinfeld, A., & Zimmerman, J. (2019, May). Unremarkable AI: Fitting Intelligent Decision Support into Critical, Clinical Decision-Making Processes. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (pp. 1–11). Glasgow Scotland Uk: ACM. Retrieved 2022-08-04, from https://dl.acm.org/doi/10.1145/3290605.3300468 doi: 10.1145/3290605.3300468

Yeung, K. (2018). A study of the implications of advanced digital technologies (including ai systems) for the concept of responsibility within a human rights framework. *MSI-AUT (2018)*, *5*.

Yildirim, N., Kass, A., Tung, T., Upton, C., Costello, D., Giusti, R., … others (2022). How experienced designers of enterprise applications engage ai as a design material. In *Proceedings of the 2022 chi conference on*

human factors in computing systems (pp. 1–13).

Yildirim, N., Oh, C., Sayar, D., Brand, K., Challa, S., Turri, V., … others (2023). Creating design resources to scaffold the ideation of ai concepts. In *Proceedings of the 2023 acm designing interactive systems conference* (pp. 2326–2346).

Yildirim, N., Pushkarna, M., Goyal, N., Wattenberg, M., & Viégas, F. (2023). Investigating how practitioners use human-ai guidelines: A case study on the people+ ai guidebook. In *Proceedings of the 2023 chi conference on human factors in computing systems* (pp. 1–13).

Yolles, M. (2021). Metacybernetics: towards a general theory of higher order cybernetics. *Systems*, *9*(2), 34.

Yu, B., Yuan, Y., Terveen, L., Wu, Z. S., Forlizzi, J., & Zhu, H. (2020). Keeping designers in the loop: Communicating inherent algorithmic trade-offs across multiple objectives. In *Proceedings of the 2020 acm designing interactive systems conference* (pp. 1245–1257).

Zhang, A., Boltz, A., Lynn, J., Wang, C.-W., & Lee, M. K. (2023). Stakeholder-centered ai design: Co-designing worker tools with gig workers through data probes. In *Proceedings of the 2023 chi conference on human factors in computing systems* (pp. 1–19).

Zhu, H., Yu, B., Halfaker, A., & Terveen, L. (2018). Value-sensitive algorithm design: Method, case study, and lessons. *Proceedings of the ACM on human-computer interaction*, *2*(CSCW), 1–23.

Zhu, J., Liapis, A., Risi, S., Bidarra, R., & Youngblood, G. M. (2018). Explainable ai for designers: A human-centered perspective on mixed-initiative co-creation. In *2018 ieee conference on computational intelligence and games (cig)* (pp. 1–8).

Zidane, Y. J.-T., & Olsson, N. O. (2017). Defining project efficiency, effectiveness and efficacy. *International Journal of Managing Projects in Business*, *10*(3), 621–641.

Zimmerman, J., & Forlizzi, J. (2014). Research through design in hci. In *Ways of knowing in hci* (pp. 167–189). Springer.

Zuckerberg, M. (2018, December). *Facebook Post* [Social Media]. Retrieved 2023-09-27, from https://www.facebook.com/zuck/posts/one-of-our-big-focus-areas-for-2018-is-making-sure-the-time-we-all-spend-on-face/10104413015393571/

Zytko, D., J. Wisniewski, P., Guha, S., PS Baumer, E., & Lee, M. K. (2022). Participatory design of ai systems: Opportunities and challenges across diverse users, relationships, and application domains. In *Chi conference on human factors in computing systems extended abstracts* (pp. 1–4).

Appendix

Chapter 4

Box 2. Wellbeing Theory

Conceptualizing Wellbeing

According to the World Health Organization, wellbeing is "a state of complete physical, mental, and social well-being, and not merely the absence of disease or infirmity" (World Health Organization, 1998). In other words, wellbeing is more than just being physically healthy—it also includes being mentally and emotionally healthy and feeling like you belong to and are supported by a community.

The academic literature consists of many ways to conceptualize and operationalize wellbeing. Some common dimensions of wellbeing include physical health, mental health, emotional health, social health, and spiritual health. While there is agreement among scholars, a strong consensus on the definition of the concept of wellbeing seems absent (Dodge, Daly, Huyton, & Sanders, 2012). Academics criticize the field on the basis that definitions are heavily dependent on the cultural background of the researcher and the application area of the research (Alexandrova, 2012). Considering the conceptual dissensus, a review by (Cooke et al., 2016) identifies four main areas of wellbeing literature which will be used as a framework in this paper.

Hedonic models of wellbeing focus on both pleasure and happiness. This field is pioneered by Ed Diener's tripartite model of subjective wellbeing (Diener et al., 1985), which considers satisfaction with life, the absence of negative emotions, and the presence of positive emotions, as vital components of wellbeing. While this perspective is often envisioned in terms of these three facets, his theoretical model is considered to be among the most influential ones and the "Life Satisfaction Score" is one of the most common instruments for measuring wellbeing in the world (Cooke et al., 2016; Linton et al., 2016). Eudaimonic models of wellbeing offer research that tries to account for more than the pleasure of a satisfied life. For example, Ryff's six-factor model of psychological wellbeing focuses on self-acceptance, positive relations with others, autonomy, environmental mastery, purpose in life, and personal growth (Ryff & Keyes, 1995; Ryff & Singer, 2006).

Further, Martin Seligman's wellbeing theory encompasses both perspectives (hedonic and eudaimonic) stating that wellbeing (or flourishing) can be conceptualized in terms of positive emotion, engagement, meaning, positive relationships, and accomplishment (PERMA) (Seligman, 2011). A third category of wellbeing research focuses on quality of life (QoL). Cooke et al. (2016) note that while this term is often used interchangeably with wellbeing, it should be seen as a separate category because research on QoL generally conceptualizes wellbeing to encompass models of physical, psychological, and social functions. It is often associated with wellbeing towards or during end of life and living with a disability. A common assessment instrument is the Quality of Life Inventory developed by Frisch, Cornell, Villanueva, and Retzlaff (1992). Lastly, Cooke et al. (2016) describe a fourth category that is called wellness. They note that wellness approaches are often rooted in counseling and tend to be broader and less clearly defined and not necessarily associated with assessment instruments. Rather, wellness practitioners focus on a holistic lifestyle that can include many areas of health and functioning including spiritual health. **Community Wellbeing**

Wellbeing is often understood as centered around individual experiences. However, wellbeing for a person is also dependent on a "set of interlocking issues and constraints and embedded in a dynamic social context." (Phillips & Wong, 2017) Musikanski et al. (2020) consider community well-being to include (1) community, (2) culture, (3) economy-standard of living (which includes housing, food, transportation and information and communication technology), (4) education, (5) environment, (6) government, (7) health, (8) psychological well-being, (9) subjective well-being and affect, (10) time balance and (11) work.

Box 3. Wellbeing Assessment

Measuring Wellbeing

In a recent review of 99 self-report assessments of wellbeing in adults (Linton et al., 2016), the authors note that there are a vast range of instruments based on different fundamental theories. In their review, they suggest that two of the most influential theories are subjective wellbeing from Diener et al. (1985) and psychological wellbeing from Ryff and Keyes (1995). They conclude that different instruments may be suitable depending upon the needs of the context. This sentiment is echoed in another recent review of 42 instruments (Cooke et al., 2016).

Despite the lack of convergence in academia, a recent McKinsey report on wellbeing in Europe states that "a consensus is nevertheless emerging

on how best to measure well-being. Researchers now tend to ask a basic question: "Overall, how satisfied are you with your life nowadays?" (Allas et al., 2020). This question, based on Diener's life satisfaction measure, is appealing because it is short and because it allows for comparison between populations and over time. However, this measure does not specifically provide information about what is wrong or what might help.

Domain-specific Wellbeing

While life satisfaction scores provide an excellent means for comparison, an assessment of wellbeing may be intended to inform useful actions to support improved wellbeing. For instance, measures of employee satisfaction are typically undertaken with the goal of improving employee satisfaction. Because of this, wellbeing assessments should ideally be sensitive to the needs of a particular domain. In a primary school setting, for instance, bullying may have a significant effect on a student's wellbeing; in a company setting, work-life balance may have a significant effect on employee wellbeing. In both cases, a domain-specific measure can be more useful for informing actions that may help improve wellbeing in the specific context.

When efforts are made to assess wellbeing in specific domains, like work or school, we refer to this as a domain-specific wellbeing assessment (e.g., Gregory et al., 2019; Renshaw & Bolognino, 2014; Renshaw et al., 2014). For instance, the College Student Subjective Wellbeing Scale (CSSWQ) has been designed to assess a combination of relevant components for college students which the researchers refer to as "covitality" Renshaw et al. (2014). The different components include Satisfaction with Academics, Academic Grit, School Connectedness, Academic Self-Efficacy and College Gratitude.

COVID-19 and Student Wellbeing

The COVID-19 pandemic and lockdowns had a significant impact on subjective wellbeing around the world (e.g., Aucejo, French, Paola, Araya, & Zafar, 2020; De Pue et al., 2021; C. Hu, Chen, & Dong, 2021; Khan, Shah, & Shah, 2021; White & Van Der Boor, 2020). Common topics studied include anxiety, loneliness, psychological stress, and post-traumatic stress disorder (Xiong et al., 2020). These effects may be especially amplified in university students as many tend to live in small housing, away from their families, and experience financial instability. Aside from that, students were expected to complete their educational goals as if it were a normal situation despite the many factors restricting them (e.g., internet connection, lack of jobs, family challenges) (D. N. Crawford, 2020). According to the literature, students suffered from decreased motivation (Tan, 2020),

hopelessness (Pretorius, 2021), and depression (Fawaz & Samaha, 2021). In response, some publications suggest that mindsets should be changed: for instance, grit and gratitude (Bono, Reil, & Hescox, 2020) or optimism (Genç & Arslan, 2021) are offered as approaches to improve wellbeing and cope with the pandemic. These recommendations, however, do not directly indicate how communities or organizations might respond to improving student wellbeing.

Box 4. Context-sensitivity, actionability, and assessment experience

Context-sensitivity

Socially disruptive events, such as a pandemic, can trigger changes in human values and their prioritization in society (Daher, Carré, Jaramillo, Olivares, & Tomicic, 2017; Klenk & Duijf, 2021). Due to isolation and lockdowns during COVID-19, there seemed to be many factors that were previously not considered as critical to wellbeing. For instance, the experience of one's home working-environment—factors such as 'Wi-Fi quality' or 'a dedicated work desk' are generally not considered by wellbeing assessment instruments. Yet, in the context of COVID-19, these factors became relevant to the wellbeing experience of community members. It is a general challenge for design research to identify the various mechanics that affect wellbeing (Fokkinga et al., 2020). Therefore, we needed a method that could identify important new factors—to identify if we were asking the right questions to the right people. This method should help identify what factors are currently actively impacting wellbeing in a manner that can point to where interventions should and can be designed.

Actionability

"Off-the-shelf" measures of wellbeing, mainly found in psychological literature (as discussed in the previous section), are oftentimes constructed primarily for validity and reliability—not actionability. What we define as actionability is the usefulness of a measure for informing helpful actions. For example, the Satisfaction with Life Scale (SWLS) (Diener, 2019) has been proven to be a strong cross-cultural measure of a person's wellbeing, but it is not designed to indicate how to improve wellbeing within a specific context. To illustrate, imagine you are an administrator aiming to improve the wellbeing of your members in your organization, knowing that the average member in your community has an SWLS score of 21, and a PANAS score of 26, does not immediately inform you on where you might take actions to

improve these scores. The scores must be related to contextual factors in order to be meaningful. On the other hand, in the domain of universities, the College Student Subjective Wellbeing Questionnaire (CSSWQ) (Renshaw et al., 2014) may be more actionable than a general measure (such as the SWLS) due to the granularity of its questions. But, despite this granularity, the questions do still not directly point to opportunities for taking action. Current measures of wellbeing may be reliable and valid and yet the information provided by these measures may not be sufficiently concrete that communities might use to take action to support improved wellbeing. Note that it is not specifically items that provide more actionable information, it is the assessment instrument as a whole, combining "off-the-shelf" measures with contextualized items. Hence the term context-sensitive assessment—not measure—of wellbeing.

Assessment Experience The assessment experience is important for two basic reasons. First, a positive experience can lead to improved participant engagement and data quality (Baumgartner et al., 2021; Stocké & Langfeldt, 2004). Second, the experience of assessing wellbeing has the potential to offer an intervention in and of itself. Namely, reviewing different facets of one's own life has the potential to lead to constructive change and experiences of improved wellbeing. While this second rationale for improving the wellbeing assessment experience was not quantitatively evaluated in this study, it was a driving motivation for the design of My Wellness Check.

Chapter 5 - Narratives for Expert Study

MiHue

Sarah Finds Her Person on MiHue

Sarah was a 24-year-old marketing assistant who had recently moved to Amsterdam for her job. Though doing well at work, Sarah hoped to expand her social circle and meet a romantic partner.

Sarah sighed as she swiped left on another dating profile. "No luck tonight?" her friend Amanda asked, noticing her frustration. "Ugh, no," Sarah replied. "I'm so over these apps only focusing on looks and generic interests. The conversations are meaningless." Amanda nodded, "It's impossible to make real connections on them."

"Exactly!" Sarah said. "I want someone I can have deep talks with, not just small talk." Then, Amanda mentioned a new app called MiHue that matched based on compatibility, not appearances. Intrigued, Sarah downloaded it, hoping to find someone she could truly connect with.

Sarah spent time customizing her profile to accurately convey herself as a person. The app first asked for basic information like her age, location, interests, and hobbies. Sarah entered details such as her love of books, yoga, piano music, cooking and indie films. There was also a section to enter personality traits and values. After thinking about it, Sarah chose words like ”kind”, ”quirky”, ”adventurous” and ”curious”. She hoped that showing these parts of her real, authentic self would help attract like-minded matches.

Next was photo selection. MiHue automatically sorted Sarah's camera roll into categories based on interests she had entered, like “Book Club”, “Yoga Poses”, and even a category for her cats! The app recommended choosing a thoughtful balance of photos, showcasing interests she shared with many others, as well as unique photos that highlighted her individuality. Following this advice, Sarah picked a mix of photos showing herself reading at book club, doing yoga poses, snuggling with her cats, dressed up silly for a costume party, and exploring street markets while traveling solo. She appreciated MiHue's guidance in thoughtfully selecting photos to give a real glimpse into her life.

Before completing her profile, MiHue generated bio suggestions based on Sarah's selected interests and personality traits. Sarah was pleased to see MiHue recommend phrases and descriptions that she identified with, like “eager world traveler” and “loves learning”. After adding this personalized text to her bio, Sarah felt confident she had shown a genuine, multi-sided portrayal of herself. She hoped this openness would attract partners interested in the same kind of real connection.

Sarah then applied filters to further customize the profiles she would see on her swiping screen. She highlighted interests and values important to her, like “book lover”, “yoga fan”, “stargazing” and “kindness”. MiHue suggested more detailed selections based on Sarah's existing choices, such as her favorite book type, yoga style, and specific constellations. Adding these helped fine-tune her results beyond surface-level interest matches. Sarah was eager to start swiping and see if these filters would find her perfect partner!

Sarah was excited to see MiHue highlight keywords and interests she had in common with each potential match as she swiped through profiles. Whenever she matched with someone, a pop-up would alert her to any especially unique interests that she and her match shared. Seeing that a match was equally passionate about an obscure fantasy novel series, or appreciated her favorite niche yoga philosophy, immediately captured Sarah's attention. It sparked a feeling of kinship, as these rare commonalities carried more weight and fostered a deeper connection. Sarah realized that even such simple similarities meant much more to her than merely finding someone attractive.

After swiping for quite some time, MiHue checked in to gather Sarah's feedback on her experience so far. Sarah noted how much she appreciated connecting based on shared values, passions, and personality traits, rather than just appearances. MiHue processed this input, and Sarah soon noticed refined highlight suggestions based on the types of profiles she responded well to. With these tweaks, her results improved drastically, saving Sarah endless swiping by discovering ideal matches sooner.

Before long, Sarah matched with David, who shared her love of books, stargazing, cooking and costume parties. MiHue immediately suggested personalized conversation starters about their favorite constellations and stargazing spots. Sarah felt relieved that the app provided these tailored opening messages, reducing the pressure and anxiety she typically felt when having to make the first move. As she and David continued chatting with MiHue's assistance, Sarah was struck by how smoothly the conversation flowed. Rather than the typical small talk she was used to, they dove into discussing childhood memories, future dreams, and the stresses of moving to a new city.

Overall, Sarah was really impressed with how the MiHue app worked. It helped her show her true self and then actually matched her with people who shared deeper compatibility, not just superficial interests. The app seemed to really 'get' her personality and what she was looking for, based on how she filled out her profile and reacted to different matches. She felt like MiHue was tailoring its recommendations just for her, suggesting people who she could build a unique connection with, instead of the usual t generic matches.

After countless disappointing dating app experiences, MiHue had given Sarah renewed hope around finding not just a partner, but someone who would value every part of who she was. Sarah looked forward to building this new meaningful connection with David, and seeing where it led organically without any pressure. She was grateful to MiHue for restoring her faith in the process of open, authentic human connection.

FoodVibe

Sascha Explores New Musical Horizons

Sascha had been using Spotify for years, but recently they felt like they were stuck in a musical rut. Playlists like Discover Weekly and Release Radar were starting to feel boring, only suggesting songs in the genres they usually listened to, like pop, indie rock, and folk. Sascha wanted to expand their musical tastes and try new types of music, but every time they tried searching Spotify's huge catalog on their own, they felt overwhelmed and went back to their musical comfort zone.

Sascha wished Spotify had a way to guide them through new genres, encouraging them to try different music while keeping the exploration manageable and curated. One day, while using the app, an ad for a new feature called "Discover More" caught their attention. The description said this interactive experience could introduce listeners to unfamiliar genres in a way tailored to their current listening habits. It promised the chance to gain new perspectives and foster personal growth through the musical journey. Intrigued and inspired, Sascha tapped the big "Let's Go!" button right away.

The app screen changed to show a map, filled with bubbles of all sizes, each representing a different music genre. Some Sascha recognized, while others sounded completely unfamiliar. In the very center pulsed their profile bubble, showing their top genres of indie pop, folk rock, and neo soul. Using the easy touch controls, Sascha zoomed out to see genres spreading across the whole map. They felt excited to explore this new world of music outside their usual tastes.

Guided by Spotify's algorithms, Sascha started moving their profile bubble toward a nearby group of genres they knew about but rarely intentionally listened to: country, folk, bluegrass, and Americana. As they approached, the app automatically made a preview playlist mixing popular songs and lesser-known tunes. The twangy vocals, fast banjo strumming, and lyrics about small towns and country life captivated Sascha immediately. They smiled, tapping the heart icon to save several songs to a new playlist appropriately called "Country Roads."

After an hour exploring those genres, Sascha was surprised they had built a country playlist with over 50 songs. It satisfied them in a way they didn't expect, making them think about lyrics exploring topics like family, faith, and rural working class life. Occasionally, thought-provoking questions from Spotify showed up on the right side of the screen, like "What emotions do you feel from this music?" and "How might these songs connect you to new people or places?" Sascha liked that these prompts helped them to reflect on how the music impacted their feelings and views.

Ready for the next part of their journey, Sascha used the touch controls again to zoom out and browse nearby areas. One cluster labeled Afrobeat, reggae/dub, soca, and dancehall caught their attention. They moved their profile bubble there, excited by the preview's lively instruments, upbeat rhythms, and chanting vocals. As the first few songs played, Sascha's shoulders started swaying instinctively to the infectious beats. The music felt vibrant, celebratory, and liberating. They soaked in information about each genre's history while listening, appreciating them in a richer context.

After a while, Sascha glanced at the map and was amazed to see how

far their profile bubble had moved from the center. Music styles they never would have tried before now characterized their soundscape. Sascha realized this journey had expanded their tastes in ways they didn't think were possible, unlocking new understandings and perspectives.

As the hours passed by quickly, Sascha felt themself growing mentally tired. But they were thrilled by all the new music worlds they had uncovered. Looking at their library, they now had playlists labeled Country Roads, Island Vibes, African Beats, and more. It was time to finish this session, but Sascha knew this was just the beginning of a lifelong musical adventure. They could return to Discover More anytime, choose a new direction, and keep growing.

Thinking about the experience, Sascha was grateful to Spotify for making Discover More. Far from just an algorithm-driven music finder, it felt like a service designed to broaden Sascha's perspectives while respecting their choices. The app had achieved its goal: personal growth through exploring music. Sascha went to bed that night feeling their world had expanded, with endless possibilities ahead.

Explore More

Andrea's Path to Mindful Eating

It was around 6pm when Andrea's stomach started growling. They needed to figure out some dinner. With a groan, Andrea went to the kitchen. Opening the fridge, they saw some vegetables, tofu, yogurt and condiments. The cabinet had a few canned goods and some pasta. Lately, their life had felt very busy with work and friends. Making dinner was often forgotten—most nights they would just order takeout food or heat up a frozen meal.

But for some reason, Andrea didn't feel like more greasy takeout tonight. They wanted something homemade and healthy. If only they had more ingredients to use...

Suddenly, Andrea remembered their friend Taylor mentioning a meal planning app called FoodVibe. "It's great! It helps me cook more carefully and be mindful about my eating habits," Taylor had said excitedly. "You have to try it, Andrea!" Well, now was a good time, Andrea thought. They downloaded FoodVibe on their phone and made a profile. For their goal, they put "eating more carefully."

Andrea found the app very easy to use. It immediately asked them to take some photos of the ingredients they had. Andrea arranged the vegetables and tofu nicely for a little photoshoot, then uploaded the pics to the app. In a matter of seconds, FoodVibe made a list of recipes they could make

using just those ingredients. One dish, a vegetable stir fry, looked good to Andrea—perfect for tonight!

As they started preparing, Andrea tried to follow the careful eating advice from FoodVibe. They focused on the colors and textures of the vegetables as they chopped...the sizzling sounds as the food hit the pan...the delicious smells filling the kitchen. Cooking this way felt calming, almost meditative. Before Andrea knew it, their stir fry was done! They quickly took a pic for the app before eating.

The meal tasted amazing–fresh, healthy, and so satisfying. Andrea felt proud that they made it themselves with just the ingredients they had. After eating, they labeled the photo in FoodVibe as "tasty" and "easy" and saved it to look at later.

Over the next few weeks, Andrea used FoodVibe daily to plan and log meals. Taking food photos and labeling recipes became a helpful routine, creating a visual record that made them appreciate and think about each meal more. Looking at their FoodVibe journal also gave Andrea some important insights. They saw that although takeout had made up most of their diet, cooking healthy meals at home gave them energy in a different way.

Using the app's features regularly helped Andrea get more organized with preparing food. They learned go-to homemade recipes they loved eating again, including that tasty vegetable stir fry. Following FoodVibe's careful eating advice improved Andrea's enjoyment of home cooking. Over time, using FoodVibe gave Andrea a real sense of achievement—they were making real progress towards their goal of developing a healthier relationship with food. The app provided helpful tools that supported their continued learning and growth around careful eating.

A few weeks later, Andrea met their friend Taylor for coffee. Andrea and Taylor had been close since college, but recently they hadn't been seeing each other that often due to their busy lives.

"Thanks for telling me about FoodVibe—I'm loving it!" Andrea said.

"So happy it's working for you as well! We'll have to get together and cook something fun from it soon." Taylor suggested.

"That's a great idea! Let's plan a dinner date." Andrea replied excitedly.

They both opened the FoodVibe app on their phones. Andrea & Taylor tapped on the "Food Friend Finder" feature. Suddenly, each of their apps detected that another user was sitting close by. Based on the overlap of recipes they had logged as having enjoyed, FoodVibe recommended a Mediterranean chickpea skillet for their dinner date.

"Ooh that looks delicious, let's make it together!" said Taylor.

Andrea smiled, excited to reconnect more with their old friend over a home-cooked FoodVibe meal. They were grateful the app could bring users together in such a tangible way.

The day of their dinner date finally arrived, and Andrea went over to Taylor's apartment, excited to cook with their friend again. In the kitchen, they scrolled through the Mediterranean chickpea skillet recipe in the FoodVibe app, splitting up the tasks.

Once the skillet was in the oven, the two friends caught up on life while sipping wine. It felt just like old times. When the timer went off, they opened the oven to reveal a beautifully aromatic dish.

Over the meal, Andrea and Taylor continued bonding over their love of food. They took pictures of the delicious chickpea skillet to log in the app later. Andrea labeled the dish as ”fun”, ”exciting”, and ”easy” in FoodVibe. They knew the app could use these labels to recommend similar fun and easy recipes to make in the future. After dinner, the pair browsed FoodVibe some more, planning more recipes to cook the next time they got together.

Andrea felt so grateful for the app bringing them and Taylor back together. They hoped they would keep using FoodVibe to explore mindful cooking and reconnect over homemade food. The app provided an easy way to share recipes, photos and memories.

In the following weeks, Andrea and Taylor met up to cook several more times. They loved learning new recipes, cooking tips and nutrition facts together through FoodVibe. Using FoodVibe became a regular ritual that strengthened their friendship, helping them form deeper bonds with each other through food.

Acknowledgements

Derek, from our very first meeting, you said, “To do this, we have to become a team.” I believe we achieved just that. Perhaps an unorthodox team, yet over the years we have developed and defended a shared vision, even when this vision was not always appreciated. You have profoundly influenced both my professional growth and personal development. I am inspired by the enthusiasm and intellect with which you engage the world every day, and I aspire to bring the same qualities to my own interactions. Someday, I hope to grasp the *vibe* as well as you do, so I can follow it wherever it leads me

Paul, it has been a true honor working with you. Your support has seen me through tough times and helped keep confidence in myself and my work. AI was initially not really your *thing.* Yet, through our countless, sometimes slightly heated, conversations about cybernetic systems and the nature of wellbeing, you have elevated the quality of this project invaluably. Your meticulous approach to research—leaving no stone unturned until every concept is clear—has been profoundly inspiring, and I will carry this forth in my academic career. I am grateful for the intellectual integrity and personal support you have provided, which have been instrumental in my development as a researcher and thinker.

To my paranymphs, Cais and Mireia, your support has been invaluable. Cais, my best bud since arriving at TUD. Bless your heart, you’re the kindest person I know. Thank you for carrying me through those earlier years and for listening to my silly impressions of every English accent—Sláinte. Mireia, I cherished our time bonding over “dos cañas” under the sun in your beautiful home country of Basque. Your energy and critical insight have guided me through academically challenging times—and I am forever thankful.

I extend my gratitude to everyone at Studio Lab—Aadjan, Evert, Iohanna, Laura, Sofie, Değer, and Nazli. Despite the solitude brought on by COVID, the moments shared with you have been among the most inspiring and comforting, truly making the Lab feel like a second home.

A special shoutout to Luce, the best desk mate one could ask for, who infused my mornings with Italian spirit and introduced me to incredible collaborative opportunities. Working with you has been a joy.

To the DIOPD team—Siyuan, Haian, Zhuochao, Alev, Pelin, Meike,

Hazal, and Pieter—thank you for your support, stimulating discussions, and nice dinners, as well as to Lisa and Anna for the enriching conference conversations.

My fellow PhDs— Timothy, Jasper, Mahan, Alec, Fredrik—it was good to see friendly faces roaming the same halls. Gijs, we still have to plan that Pubquiz! And, Vera, who would have thought we would be drinking Old Fashioneds together in the US of A when we were doing our bachelors?

I am immensely grateful to my graduation students—Céline, Garoa, Emma, Ziyi, Chia-Ling, Henrique, Joseph, Irmak, Moshiur, Bill, Neel, Markus, Charlotte, and Mark—for entrusting me with their projects. In particular, I would like to thank Yi, Danielle, and Gigi, because without you the final study in this dissertation would not have been possible.

A heartfelt thanks to the Study Climate team—Stella, Young-Mi, and Janne—for their collaboration, trust, and the inclusion in the wellbeing events, which were both enjoyable and enriching.

And of course a shoutout to *ChatGPT & Claude* for being the best assistant a PhDer could ever want.

My friends, thank you to everyone who has supported me in any way over the years, helped me see this through, listened to my ramblings, and made me into the thirty-something I am: Pim, Timon, Anna, Dickie, Kim, Ashwin, Amber, Sjors, Willem, Sem, Simon, Wuut, Marilon, Mitchie, Jasmine. But of course, also specifically the foundation and co: Bart & Parel, may the topics of our conversations and your admirable joy in life never fade, and Vonk & Renske, you are the most genuinely warm-hearted people I know, and Emiel & Suda, we better catch 'm all.

Dear *pap en mam*, your unconditional support has given me the strength and confidence needed to complete this journey. Knowing that you will always stand beside me gave me the confidence to take this monster down—thank you. And Marieke, through this project, I feel that I have come closer to the passion with which you face the world. I admire you in that and hope it never changes.

Finally, *mijn allerliefste Chanelle—mijn stralende zonnebloem.* From the cover art and diagrams to being my daily sounding board for all my ideas, your fingerprint is unmistakably on this work. Your presence and partnership have sustained me through countless challenges. Without you and Cato, I could have never done this. You are my sanctuary, the calm in every storm. As long as we are together, I know everything will be alright.

Curriculum Vitæ

Willem Lennert Antoon van der Maden

Willem van der Maden was born on August 1, 1993, in Boxmeer, The Netherlands. His academic journey began with secondary education in Nijmegen, followed by a foundational year in Industrial Product Design. Willem then moved to Utrecht University for a Bachelor of Science in Liberal Arts and Sciences, where completed dual majors in Cognitive and Neurobiological Psychology and Philosophy of Aesthetics. His interest in the complexities of human emotion, consciousness, and behavior, led him to pursue a Master of Science in Cognitive Psychology at the same university, graduating *cum laude*.

Willem's master's thesis, "What Good Are Twenty-five Positive Emotions? An Extension Of The Emotion Typology," completed in collaboration with *Emotion Studio*, studied and verified the differentiation of 25 positive emotions. This work laid the foundation for his PhD in Design at the Technical University of Delft, focusing on designing artificial intelligence that promotes human wellbeing. Here, he developed a digital wellbeing platform for wellbeing assessment used by over 20,000 members of the TU Delft community during the COVID-19 pandemic, contributing to institutional policy through community-led design and data analysis. Insights from this project also inspired the development and evaluation of a human-centered AI design method, aimed at actively fostering human flourishing.

An important part of Willem's role at TU Delft involved supervising 20 graduation projects, with 10 in the bachelor's program and 10 in the master's, all centered around design and AI. This mentorship showcased his commitment to guiding the next generation of designers and technologists. Willem's research on positive AI led to first-author publications and presentations at international conferences, further contributing to the field of human-centered AI. He also organized an impactful academic workshop on the role of emergent Generative AI technologies in the field of design practice and design research—leading to an ACM Interaction publication

and sparking the onset of an international community for collaborating on future work related to Generative AI in design.

Outside of academia, Willem appeared on a national Dutch news channel (RTL) to discuss generative AI and was a member of the IEEE 7020 standards review committee. Additionally, Willem played a crucial role in the TU Delft Study Climate taskforce, not only supporting the wellbeing of staff and students during the pandemic but also contributing to early government-led experiments investigating the wellbeing effects of COVID-19 in public spaces.

Currently, Willem has embarked on a new chapter as he begins a postdoctoral position titled 'Designing Human-AI Interaction for Wellbeing and Health' with Associate Professor Jichen Zhu at the Digital Design department, ITU Copenhagen, furthering his commitment to designing *Positive AI.*

List of Publications

Included publications

1. **van der Maden, W.**, Lomas, D., and Hekkert, P. (2022) Design for wellbeing during Covid-19: A cybernetic perspective on data feedback loops in complex socIotechnical systems, in Lockton, D., Lenzi, S., Hekkert, P., Oak, A., Sádaba, J., Lloyd, P. (eds.), DRS2022: Bilbao, 25 June - 3 July, Bilbao, Spain. https://doi.org/10.21606/drs.2022.771

2. **van der Maden, W.**, Lomas, D., & Hekkert, P. (2023). A framework for designing AI systems that support community wellbeing. *Frontiers in Psychology*, 13, 7900. https://doi.org/10.3389/fpsyg.2022.1011883

3. **van der Maden, W.**, Lomas, D., Sadek, M., & Hekkert, P. (2023). Positive AI: Key Challenges for Designing Artificial Intelligence for Wellbeing. *[Manuscript unver review for publication at She Ji]*

4. **van der Maden, W.**, Lomas, D., & Hekkert, P. (2024). Developing and Evaluating a Design Method for Positive Artificial Intelligence. *[Manuscript under review for publication at AI EDAM]*

Ancillary publications

1. **van der Maden, W.,** Van Beek, E., Nicenboim, I., Van Der Burg, V., Kun, P., Lomas, D., & Kang, E. (2023, July). Towards a Design (Research) Framework with Generative AI. In *DIS '23 Companion: Companion Publication of the 2023 ACM Designing Interactive Systems Conference* (pp. 107-109). https://doi.org/10.1145/3563703.3591453

2. Lomas D, Lin A, Dikker S, Forster D, Lupetti ML, Huisman G, Habekost J, Beardow C, Pandey P, Ahmad N, Miyapuram K, Mullen T, Cooper P, **van der Maden, W.**, Cross ES. (2022) Resonance as a Design Strategy for AI and Social. *Robots. Front Neurorobot.* 16:850489. doi: https://doi.org/10.3389/fnbot.2022.850489

3. Lomas, J.D., **van der Maden, W.**, Lion, G., Bandyopadhyay, S., Litowsky, Y., Xue, H., & Desmet, P. (2023). Emotional Alignment of AI and Humans: Human Ratings of Emotions Expressed by GPT-3, DALL-E and Stable Diffusion. *Frontiers in Computer Science*. [*Accept with revisions]*

4. Moilanen, J., van Berkel, N., Visuri, A., Gadiraju, U., **van der Maden, W**., & Hosio, S. (2023). Supporting mental health self-care discovery through a chatbot. *Frontiers in Digital Health*, 5. https://doi.org/10.3389/fdgth.2023.1034724

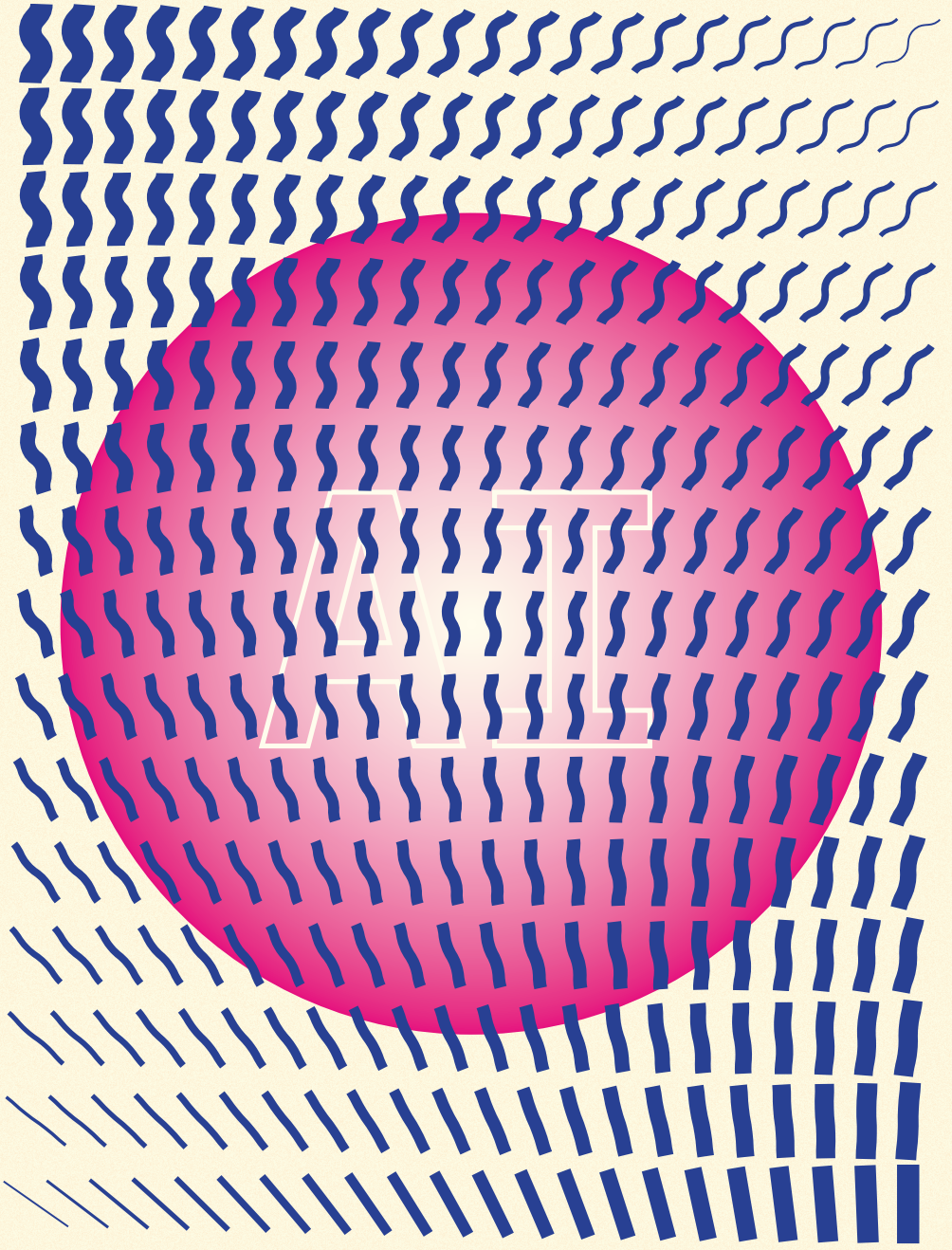
5. Lomas, J. D., & **van der Maden, W**. (2021). MyWellnessCheck: Designing a student and staff wellbeing feedback loop to inform university policy and governance. In M. van der Bijl-Brouwer (Ed.), Proceedings of Relating Systems Thinking and Design (RSDX) Symposium. (pp. 247-260) https://rsdsymposium.org/mywellnesscheck-designing-a-student-and-staff-wellbeing-feedback-loop/

6. Beardow, C., **van der Maden, W**., & Lomas, J. (2020). Designing Smart Systems: Reframing Artificial Intelligence for Human-Centered Designers. *In TMCE 2020: 13th International Tools and Methods of Competitive Engineering Symposium* (pp. 143-154).

7. Hoggenmueller, M., Lupetti, M. L., **van der Maden, W**., & Grace, K. (2023, March). Creative AI for HRI Design Explorations. In Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction (pp. 40-50). https://doi.org/10.1145/3568294.3580035

8. Hagens, E., Lupetti, M., **van der Maden, W**., Steegers, R., Rousain, M. (2023) Trustworthy Embodied Conversational Agents for Healthcare: A Design Exploration of Embodied Conversational Agents for the periconception period at Erasmus MC.In ACM conference on Conversational User Interfaces (CUI '23), July 19–21, 2023, Eindhoven, Netherlands. ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3571884.3597128

9. Gomez Beldarrain, G., **van der Maden, W. L. A**., Huang, S., & Kim, E. Y. (2023). Identifying meaningful user experiences with autonomous products: a case study in fundamental user needs in fully autonomous vehicles. https://doi.org/10.21606/iasdr.2023.434


ART