



UNIVERSITY OF
CAMBRIDGE

Designing Meaningful Algorithmic System Transparency for Non-Expert Users

Bianca Giulia Sarah SCHOR



Trinity College

Dissertation submitted in June 2024 for the degree of Doctor of Philosophy in Computer
Science and Technology

Declaration

This thesis is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the preface and specified in the text. It is not substantially the same as any work that has already been submitted, or, is being concurrently submitted, for any degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the preface and specified in the text. It does not exceed the prescribed word limit for the relevant Degree Committee.

Bianca Giulia Sarah SCHOR

June, 2024

Abstract

Transparency matters: people need it to make informed decisions regarding technologies that might impact them. Among others, non-expert users represent key stakeholders that can benefit from but also be impacted by algorithmic systems. Their transparency needs deserve urgent attention as daily lives increasingly rely on such systems. In this thesis, I explore how to design meaningful algorithmic system transparency for non-expert users.

I highlight three core barriers to transparency, among other challenges raised in the literature. First, it is unclear *how* to design transparency in practice (‘Design Challenge’). Transparency is considered a core principle of Responsible AI, but not a panacea. To serve users, I thus argue transparency must be *meaningful* to them, i.e. contextually appropriate. Secondly, new laws and transparency guidelines are emerging. I argue these must align with designers’ and users’ perspectives to become effective (‘Guideline Challenge’). Lastly, there remains the need to assess transparency (‘Assessment Challenge’).

By directly engaging with users and designers of AI systems, I provide empirical and critical insights to help address these three challenges. I thus contribute towards bridging the gap between theory and design practices, so that algorithmic systems better cater for users’ information needs in real-world settings. This is key to help ensure that systems can be reviewed, challenged, and used safely by users. Without meaningful transparency, it would not be possible to understand, critique, and scrutinise them.

To explore the Design Challenge, I first investigate what meaningful transparency means for under-studied users, in particular parents in the UK in the wake of the new online safety legislation (Chapter 3). To help solve the Guideline Challenge, I study the gap between a recent international standard on algorithmic system transparency and designers’ practices (Chapter 4). I also use a co-design approach to test and expand design recommendations from the human-centred explainable AI (HCXAI) literature in a specific clinical domain: gynaecology (Chapter 5). Lastly, I explore the Assessment Challenge by benchmarking meaningful transparency (Chapter 3) and experimenting with a critical method to investigate the transparency dimensions of a system’s user interface (Chapter 6). This thesis thus aims to provide guidance for designers, standards bodies, academics, and policy-makers on what meaningful transparency can mean for various non-expert users, and how to promote it in practice.

Acknowledgements

This PhD thesis would not have been possible without the endless support and precious help from my two supervisors Prof Alan F. Blackwell and Dr. Jatinder Singh. Thank you very much to both of you for giving me the opportunity to carry out this research project, and for being so encouraging from start to finish. I am also incredibly grateful to my fabulous friends, colleagues, favourite housemates, and wonderful siblings who not only helped me complete my PhD degree but also made these the best years of my life. I am especially thankful to the Cambridge University Triathlon Club and its inspiring members for helping me improve my health and challenge myself everyday. I particularly thank Chris Norval, Michelle Seng Ah Lee, Ed Cator, and Ellen Matin Charlesworth for helping me proofreading my thesis. Lastly, I would like to dedicate this thesis to my grandmothers, Gretta Schor and Anna Ristori, who are incredibly strong, caring, and inspiring women.

Contents

1	Introduction	21
1.1	Publications during the PhD	21
1.2	Goal overview: promoting technologies that effectively serve people	22
1.3	Problem description and main research question	23
1.4	Three core challenges to building transparent algorithmic systems	23
1.4.1	RQ1: The Design Challenge	24
1.4.2	RQ2: The Guideline Challenge	24
1.4.3	RQ3: The Assessment Challenge	25
1.5	Thesis outline and main contributions	25
1.6	Defining algorithmic system transparency, non-expert users, and design . .	28
1.7	Reconceptualising meaningful transparency	30
1.8	Positionality	31
1.9	Ethical approval	32
2	Literature review	33
2.1	Defining and justifying transparency for users	33
2.1.1	Opaque systems and associated risks for users	33
2.1.2	Transparency goals at a high level	35
2.1.3	Transparency goals in context	36
2.1.4	Transparency as a means to promote accountability	37
2.1.5	Defining transparency, explainability and interpretability	38
2.2	The Difficulty of improving transparency in practice	39
2.2.1	The Principles-to-practice gap	39
2.2.2	Towards a socio-technical system approach	40
2.2.3	Better defining non-expert users and their information needs (Literature gap 1)	41
2.2.3.1	HCTXAI: Recentring transparency research on non-expert users	42
2.2.3.2	Analysing algorithmic system transparency in user interfaces	43

2.2.4	Exploring the gap between guidelines and design practices and operationalising HCXAI recommendations in context (Literature gaps 2 & 3)	44
2.2.4.1	Transparency tools, standards and designers' approaches	44
2.2.4.2	Operationalising HCXAI recommendations	45
2.2.5	Can we measure transparency? (Literature gap 4)	45
2.3	Addressing transparency risks and trade-offs	46
2.3.1	Transparency vs Accuracy: a key trade-off to explore?	46
2.3.2	Transparency vs Privacy: a complex relationship	47
2.3.3	Transparency fallacy & the transparency paradox	48
2.3.4	Towards meaningful transparency in context	48
2.4	Summary and main takeaways	49
3	Understanding users' information needs: parents in the UK	51
3.1	Introduction	51
3.1.1	Why focus on parents in the UK?	52
3.1.2	Promoting meaningful transparency for digital parenting	52
3.1.3	Chapter outline and contributions	53
3.2	Theory: Defining meaningful transparency for digital parenting	54
3.2.1	Transparency for parents and its limitations	54
3.2.2	Contextual integrity as a framework to define meaningful transparency for digital parenting	55
3.3	Background: how meaningful transparency can support digital parenting	56
3.3.1	Meaningful transparency for whom? Parents' role in children's online safety & privacy	56
3.3.2	Parents need meaningful information to address their legal duties and the risks children face online	57
3.3.3	Early teenagers are particularly vulnerable so parenting needs to adapt to oversee new technologies	58
3.4	Method	59
3.4.1	Parents and their communities	59
3.4.2	Demographic information and ethical considerations	60
3.4.3	Interview method and collaborative drawing activity	60
3.4.4	Data analysis	61
3.5	Key insights from online interviews with parents	63
3.5.1	Participants know little about the technology that children use but are keen to understand them better	64
3.5.2	Missing information in participants' current approach to digital parenting	64

3.5.3	Several participants feel overwhelmed with the information they receive regarding their children	65
3.5.4	Communities provide most information on technology children use, seemingly not service providers	66
3.5.5	Participants find the information they receive from fellow parents helpful	67
3.6	Discussion	68
3.6.1	Parents vs service providers' responsibility towards children's online safety	68
3.6.2	Communities as a starting point for parent-supportive design: norms of flow among fellow parents	69
3.6.3	Study limitations	69
3.6.3.1	Limitations related to our methodology.	69
3.6.3.2	Limitations related to transparency.	70
3.7	Conclusion and main takeaways	71
4	Mind the gap: transparency guidelines & designers' practices	73
4.1	Introduction	73
4.1.1	A 'Principles-to-practices' gap	74
4.1.2	Study outline and contributions	74
4.2	Background	75
4.2.1	Implementing transparency for users: the role of designers	75
4.2.2	Standards for human-centred design and transparent algorithmic systems	76
4.2.3	IEEE 7001 as an exemplar standard on system transparency	77
4.2.4	IEEE 7001's "measurable, testable levels of transparency" for users (TLs)	78
4.3	Methods: Exploring Standards & Designers' approaches to System Transparency	79
4.3.1	Mixed-method study design	79
4.3.2	Participant recruitment: designers as one group among IEEE 7001's target audience	79
4.3.3	Study one (online survey): designers' understanding and experiences of system transparency	81
4.3.4	Study two (interviews): exploring designers' opinions on transparency levels for users in context	81
4.3.5	Data analysis	82
4.4	Results	83
4.4.1	Highlights from the findings	83

4.4.2	Few participants know of or use standards to design user-facing algorithmic systems	84
4.4.3	IEEE 7001 transparency recommendations are critiqued by participants	85
4.4.3.1	Few participants say they are likely to implement the TLs.	85
4.4.3.2	IEEE 7001’s scenarios and TLs are overall clear but not precise enough to enable actual implementation.	85
4.4.3.3	Participants seem to find TLs difficult to implement and not always appropriate in the IEEE 7001 scenarios.	86
4.4.4	Key challenges for improving system transparency	88
4.4.4.1	What ‘transparency’ actually means is unclear to most participants.	88
4.4.4.2	Participants’ awareness about the risks posed by algorithmic systems seems primarily focused on privacy.	89
4.4.4.3	Participants identify several challenges to design transparency: clash of priorities, lack of resources, and clear definitions.	89
4.4.5	Support mechanisms for designers: raising awareness, providing tools, and outlining best practices	90
4.4.5.1	Raising designers’ awareness is needed.	90
4.4.5.2	Several designers identify help that could support them in designing for transparency.	91
4.4.5.3	Participants have suggestions for tools that they think can assist in making systems more transparent.	91
4.4.6	Principles for ‘good’ transparency in algorithmic systems for users .	92
4.5	Discussion	93
4.5.1	Suggestions for supporting designers in making algorithmic systems more transparent for users	93
4.5.1.1	Raising awareness about transparency among designers. .	94
4.5.1.2	Further identifying examples for future standards and transparency metrics.	94
4.5.1.3	Supporting designers in experimenting with transparency interventions.	95
4.5.1.4	Engaging various stakeholders more in standards development.	95
4.5.2	Study limitations	96
4.6	Conclusion and main takeaways	97
5	Operationalising HCXAI recommendations: a co-design approach with clinicians	99

5.1	Introduction	99
5.1.1	Transparency for clinical AI implementation	100
5.1.2	Gynaecologists as users of clinical AI systems	100
5.1.3	Study outline and main contributions	101
5.2	Background	101
5.2.1	Defining algorithmic system transparency for clinicians	101
5.2.2	AI in gynaecological imagery: an under-researched application domain in the XAI literature	102
5.2.3	HCXAI literature and co-design with clinicians	103
5.3	Methods: Co-design workshop and interviews with gynaecologists	103
5.3.1	The HCXAI literature reviews tested and extended in the co-design activities.	104
5.3.2	Case study: Deep Learning segmentation of ovarian cancer in CT scans	105
5.3.3	In-person co-design workshop with gynaecologists in Amsterdam UMC, the Netherlands	107
5.3.4	In-person interviews with gynaecologists in the UK and the Netherlands	108
5.3.5	Gynaecologists' demographics	109
5.3.6	Ethical considerations	109
5.3.7	Data analysis	110
5.4	Results	110
5.4.1	Main themes raised by participants	111
5.4.2	Axis one (Users): differentiating the information needs of different gynaecologist user groups	111
5.4.3	Axis two (Relevance): providing information in a prioritised manner in terms of perceived relevance for clinicians	112
5.4.4	Axis three (Timeline): prioritising information about the AI-based tool's conceptualisation, role, and purpose over (post-hoc) explanations	113
5.5	Discussion	114
5.5.1	Design recommendations for meaningfully transparent AI systems for clinicians	114
5.5.1.1	Prioritising information categories about the why, evaluation, personalisation, and integration into clinical practice.	114
5.5.1.2	Providing prioritised information in terms of timeliness and relevance for clinicians throughout the AI design lifecycle.	115
5.5.1.3	Accounting for multidisciplinary teams (MDT) using a clinical AI system by taking into account the perspective, role, and human interactions of various types of gynaecologists.	116

5.5.1.4	Providing transparent information on the AI-based tool's conceptualisation, role, and purpose prior to its implementation.	116
5.5.2	Women's health in the HCI and AI literature: engaging directly with gynaecologist users	116
5.5.3	Study limitations	117
5.5.4	Further research towards implementing meaningfully transparent AI systems for clinicians	118
5.6	Conclusion and main takeaways	118

6 Assessing transparency: a critical method to account for the socio-technical context 121

6.1	Introduction	121
6.1.1	A critical and socio-technical system approach to assess transparency	122
6.1.2	Chapter outline	123
6.2	Background: applying object-based analysis to HCI	123
6.3	Object-based analysis methods: accounting for the visual rhetoric of transparency	125
6.3.1	Object-based analysis as a methodology to study interface design .	125
6.3.2	Object-based analysis helps to better account for the visual rhetoric around transparency	126
6.3.2.1	The Capability approach applied to algorithmic system transparency.	126
6.3.2.2	Object-based analysis helps to assess how interface design influences the capabilities related to transparency.	128
6.4	APB and Parcoursup: historical background	129
6.4.1	The Introduction of algorithms in French university admissions . . .	130
6.4.2	Political drivers: design for transparency?	131
6.5	Visual analysis: a tool to assess System transparency in user interfaces . .	133
6.5.1	Logos: failing to align with users' visual culture can hinder their understanding of the system	134
6.5.2	Login pages: visualising constraint	136
6.5.3	Admissions pages: allocation transparency	140
6.5.4	Preliminary conclusions and Parcoursup's public reception	142
6.6	Discussion	143
6.6.1	Future work: integrating object-based analysis into design practice .	145
6.7	Conclusion and main takeaways	146

7	Discussion	149
7.1	Addressing the four literature gaps	149
7.1.1	Literature Gap 1: Better defining the needs of non-expert users . .	149
7.1.2	Literature Gap 2: Exploring the gap between new transparency guidelines and design practices	150
7.1.3	Literature Gap 3: Operationalising HCXAI design recommendations in real-world settings	150
7.1.4	Literature Gap 4: Evaluating transparency in practice	151
7.2	Summary of main takeaways	151
7.2.1	New conceptualisation of meaningful transparency for non-expert users	152
7.3	Other contributions to HCI and AI research	153
7.4	Main limitations of the thesis	154
7.5	Further research on meaningful transparency	155
8	Conclusion	157
	Bibliography	161
A	Chapter 3 Appendix	199
A.1	Interview script	199
A.1.1	Introduction.	199
A.1.2	General experience, concerns, and information needs regarding social media and online games used by children.	199
A.1.3	Drawing exercise.	200
A.1.4	Demographics.	201
A.1.5	Feedback and snowball sampling.	201
B	Chapter 4 Appendix	203
B.1	Glossary	203
B.2	Scenarios	204
B.2.1	<i>Moderation</i> : Content Moderation for AI.	204
B.2.2	Scenario <i>Credit</i> : Credit Scoring System.	204
B.3	Summary of Demographics from our survey (Study 1)	205

Glossary

Accountability “As a mechanism, it is an institutional arrangement whereby an actor provides accounts to a forum, who deliberates on those accounts and may impose consequences” [42]. A developer might provide information to a regulator about their system, for example, with the regulator then issuing a penalty or requiring design changes (...) By contrast, accountability as a virtue is a normative concept, a set of standards for evaluating behaviour—often tied to being transparent, responsible, and responsive—with ‘being accountable’ seen as a positive quality of particular actors [43]” [70]. For example, accountability could be a designer providing local authorities with information about the ‘smart’ toaster they have built for their neighbour and the induced risks of fire hazard.

Algorithm A computational method, formula, or procedure [203]. In daily life, a simple algorithm could be: if I put a slice of bread in the toaster, turn the power on, and wait 1 minute, then I get a piece of toast. All AI, ML and autonomous systems use algorithms.

Algorithmic system transparency There is considerable ambiguity and variations about what transparency means [15]. I here define it based on one of my key sources, the IEEE 7001 ‘Standard for Transparency of Autonomous Systems’ [171, 364] as a transfer of information from an algorithmic system or its designers to a stakeholder that is truthful; contains information relevant to the causes of some action, decision, or behaviour; and is presented at a level of abstraction and in a form meaningful to the stakeholder. Transparency should be mindful of the stakeholders’ likely perception and comprehension, and should avoid disclosing information in a manner that, while technically true, is framed in a way that leads to misapprehension. For example, transparency could be the transfer of information from a pair of autonomous trousers (similar to the one in the Wallace and Gromit film “The Wrong Trousers” [273]) or its designers to the person wearing them about the trousers’ behaviour, e.g. how the trousers have been designed or why the trousers choose a certain route, in a way that the wearer can understand and find relevant.

Algorithmic systems Following Van de Poel, I define algorithmic systems in this thesis

as socio-technical systems “that are composed of technical artifacts, human agents, and institutions [and]—in addition—contain artificial agents and certain technical norms that regulate interactions between artificial agents and other elements of the system” [346]..

Artificial intelligence (AI) AI is an ambiguous and overloaded term [203]. There is no consensus on what algorithms AI encompasses [186]. I follow Russell and Norvig’s definition: “the designing and building of intelligent agents that receive percepts from the environment and take actions that affect that environment” [295]. This could be for example a ‘smart’ toaster machine that automatically toasts bread every morning at 7 o’clock, similar to the automated breakfast robot in Wallace and Gromit films [273]. Although there is no consensus on what AI means, it includes autonomous systems, ML, and some CAD systems.

Automated decision-making (ADM) A decision-making process that involves a substantial automated component by technological means. Solely ADM has no human involvement [76] to be distinguished from non-solely (or partial) ADM [204]. ADM is what makes a system autonomous: a system that has the capacity to make decisions itself in response to some input data or stimulus with a varying degree of human oversight or intervention depending on the system’s level of autonomy [171, 364]. This could be for example a pair of trousers that make you walk automatically without any effort nor button to press, and that chooses the route you take without you giving directions, like in the Wallace and Gromit film [273].

Black-boxes AI systems that are opaque (i.e. not transparent) and whose behaviour and actions are difficult to explain/interpret, even by experts. This is often due to the complexity of the algorithms used.

Clinical AI AI applied to or relevant to the clinical domain. Clinicians include, but are not limited to, doctors, nurses and other professions allied to health [317].

Co-design “[The] creativity of designers and people not trained in design working together in the design development process” [299], for instance Gromit working with Wallace to design a personalised ‘smart’ toaster for him [273].

Computer-Aided Diagnostic (CAD) systems Systems that assist medical professionals in the interpretation of medical images during the diagnosis process [101, 104]. Most of these systems use AI.

Design I describe here the design of an algorithmic system, not design in general. I define it as the end-to-end process and final system resulting from conceiving, training,

testing, and maintaining of an algorithmic model, as well as its interface and potential hardware (e.g. ‘smart’ toaster).

Designers Designers, developers, builders, maintainers, and operators, as well as decision-makers and procurers in organisations using and deploying algorithmic systems (collectively, ‘designers’) of algorithmic systems who either wish to or are required to engineer systems that have a certain degree of transparency [171, 364]. In other words, the professionals who are in charge of producing and controlling algorithmic systems and their transparency dimensions, e.g. the ‘smart’ toaster or autonomous trousers.

Explainable AI (XAI) A field of AI research that aims to provide human-understandable information for the system’s behaviour and processes [112]. This can include various types of explanations, such as an explanation about autonomous trousers’ usual behaviour (general explanation), or an explanation of a specific outcome reached by an AI system, after its occurrence (post-hoc explanation). This can be for instance explaining why a ‘smart’ toaster burnt your toasts this morning. Transparency and Explainability are linked yet different concepts. Several academics argue “it takes more than just algorithmic transparency to make [AI systems] explainable” [111].

Human-centred Explainable AI (HCXAI) A sub-field of Explainable AI (XAI) research that emphasises ‘the who’ we open black-boxes for [112]: HCXAI puts the human at the centre of technology design and advocates for a reflective socio-technical approach [107]. This means for example identifying and describing the information needs of the person wearing autonomous trousers (like in the Wallace and Gromit film [273]) about how the trousers behave.

Interpretability There is no consensus on what interpretability means. I thus provide three definitions to illustrate how recent publications have tried to outline this concept: i) “a property of machine-learned models that dictates the degree to which a human user—AI expert or non-expert user—can come to conclusions about the performance of the model given specific inputs” [107]; ii) “Interpretation is something one does to an explanation with the aim of producing another, more understandable, explanation. As with explanation, there are various concepts and methods involved in interpretation: *Total* or *Partial*, *Global* or *Local*, and *Approximative* or *Isomorphic*.” [116]; iii) “Interpretability is the ability to provide *explanations* in *understandable terms* to a human, where 1) **Explanations**, ideally, should be *logical decision rules* (if-then rules) or can be transformed to logical rules (...) 2) **Understandable terms** should be from the *domain knowledge* related to the task (or common knowledge according to the task).” [383].

Machine Learning (ML) A set of techniques used in AI to detect and extrapolate patterns from data [203, 295], including supervised learning, semi-supervised learning, unsupervised learning, and reinforcement learning. ML is used to adapt to new circumstances and to detect and extrapolate patterns [204].

Meaningful transparency A transfer of information to a specific stakeholder group that is *contextually appropriate*, i.e relevant, accurate, proportionate, and comprehensible to them [68, 252, 247].

Non-expert users (users) By non-expert users, I mean non-expert users in AI, e.g. clinicians are medical experts, but (often) non-expert users in relation to algorithmic systems. I define such users as “[people] who have only a brief interaction or who interact every day with an autonomous system” [171, 364]. Non-expert users have not received formal training to use a given system nor do they necessarily have a technical background, e.g in data science or computer science. I also include here individuals directly or indirectly impacted by such interactions, e.g. parents of teenage social media users. .

Responsible AI “[A] manifold concept that imposes the systematic adoption of several AI principles for AI models to be of practical use. In addition to explainability, the guidelines behind Responsible AI establish that fairness, accountability and privacy should also be considered” [18]. Others also include values like human control, or professional responsibility [122].

Standard “[Documented] agreements containing technical guidelines to ensure that materials, products, processes, representations, and services are fit for their purpose.” [9] In computer science, many standards are issued by official standards bodies such as The Institute of Electrical and Electronics Engineers Standards Association (IEEE SA), American National Standards Institute (ANSI) etc.

System A set of interacting data, algorithm(s), and/or model(s) to form a technical workflow or product, e.g. a facial recognition algorithm that triggers an identity verification model [203, 204].

Chapter 1

Introduction

1.1 Publications during the PhD

Here is a list of the papers I have published and/or submitted in top-tier ACM conferences and journals in Computer Science. These papers constitute the core of my PhD research, as described in Chapters 3, 4, 5 and 6.

- **Bianca G. S. Schor**, Alan Blackwell, Jatinder Singh. 2024. Towards Meaningful Transparency for Digital Parenting. In Journal of Responsible Computing (under review)
- **Bianca G. Schor**, Jatinder Singh, and Alan Blackwell. 2024. The Visual Culture of Design for Algorithmic System Transparency. An Analytic Case Study of the APB and Parcoursup Systems. In AI & Society (under review)
- **Bianca G. S. Schor**, Emma Kallina, Jatinder Singh, Alan Blackwell. 2024. Meaningful Transparency for Clinicians: Operationalising HCXAI Research with Gynaecologists. In Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24), June 3–6, 2024, Rio de Janeiro, Brazil. ACM, New York, NY, USA, 14 pages. <https://doi.org/10.1145/3630106.3658971>
- **Bianca G. S. Schor**, Chris Norval, Ellen Matin Charlesworth, Jatinder Singh. 2024. Mind The Gap: Designers and Standards on Algorithmic System Transparency for Users. In Proceedings of the ACM CHI Conference on Human Factors in Computing Systems (CHI '24), May 11–16, 2024, Honolulu, HI, USA. ACM, New York, NY, USA, 16 pages. <https://doi.org/10.1145/3613904.3642531>
- **Bianca G. Schor**, Jatinder Singh, and Alan Blackwell. 2022. Making Algorithmic Systems More Meaningfully Transparent for Users Through Visualisation. In Designing Interactive Systems Conference (DIS '22 Companion). Association for Computing Machinery, New York, NY, USA, 18–20. <https://doi.org/10.1145/3532107.3532875>

1.2 Goal overview: promoting technologies that effectively serve people

Today, algorithmic systems are used by a growing number of people in a wide range of domains. From children chatting online with friends to clinicians diagnosing patients with cancer, many of our daily decisions are supported and/or mediated by such technologies. The underlying goal of this PhD thesis is to help promote technologies that can effectively serve the people who use them and/or are impacted by them.

To be able to make informed decisions regarding technologies that might impact them, people need to receive information about the relevant systems’ actions, design, risks, and behaviours [171, 364]. I broadly define this transfer of information as a working definition of *algorithmic system transparency*. Transparency matters, especially since our daily lives are increasingly relying on algorithmic systems [320], e.g. in education and healthcare [14, 156]. Among various stakeholders –such as regulators, designers, and policy-makers– users can benefit from and be impacted by such systems. Indeed, recent technical improvements have introduced new opportunities using algorithms, such as improving patient outcomes or accelerating drug discovery [28, 311, 350]. However, algorithmic systems can also pose serious risks to individuals and communities, including privacy risks of surveillance and legal risks of discriminatory bias [265, 351]. It is thus crucial to empower users to scrutinise and challenge algorithmic systems, whenever needed [68].

Indeed, algorithmic system transparency is key to support accountability [93, 195, 250]. By making it possible to discover how and why a system made a particular decision [364] or disclosing the processes that relate to its functioning [319], transparency can help identify, investigate, and address potential biases associated with a system for instance. Importantly, any serious issue associated with an algorithmic system should lead to appropriate measures and consequences, e.g. by changing the system design or compensating those impacted. Transparency is therefore necessary for regulators and governing bodies, such as the Information Commissioner’s Office (ICO) in the United Kingdom, to ensure algorithmic systems are fit for people to use. However, transparency is also needed at an individual or community level to enable users, including those who might not have been trained to review or understand such systems (hereafter called “non-expert users”), to hold service providers and designers to account for the risks algorithmic systems pose (see §2.1).

Indeed, designers play a key role –among other stakeholders– in making sure non-expert users receive information about the systems that might impact them [99]. In line with one of my main sources, the international IEEE 7001 ‘Standard for Transparency of Autonomous Systems’ [171], I broadly define designers as people in organisations involved in the design, development, and deployment of algorithmic systems. Despite the growing calls for transparency [182], several algorithmic systems have recently led to public outcry

due to their lack of transparency. In 2021 for example, the Facebook files scandal shed light on the lack of transparency in widely used algorithmic systems such as Instagram about their impact on a broad range of users, including vulnerable users (e.g. teenagers) [133]. It is thus urgent to redress this systemic, asymmetrical power dynamic between designers and users [70], and transparency is essential towards this [247, 252].

1.3 Problem description and main research question

Transparency is considered necessary by many academics [106, 66], policy-makers [182], and regulators [78, 352]. Nonetheless, building transparent algorithmic systems remains an open challenge today [251]. Responsible AI (RAI) and HCI researchers have described a ‘principles-to-practice’ gap, i.e. a significant difference between theory and practice when it comes to RAI principles such as transparency – see §2.2.1 for a more details covering of the related literature. As mentioned above, a lack of transparency in algorithmic systems can impact various stakeholders including non-expert users. This thesis therefore explores the principles-to-practice gap by focusing specifically on non-expert users’ transparency needs. Towards this, I investigate the following Core Research Question (CRQ):

- **CRQ: How to design meaningful algorithmic system transparency for non-expert users?**

To explore CRQ, I take a human-centred, empirical, socio-technical system and impact-driven approach (see §1.7): given the challenges and limitations associated with designing algorithmic system transparency in practice (see §2.2), I argue that we need to ensure the information transferred to non-expert users is meaningful to them, i.e. *contextually appropriate* [69] (see further details in §1.4.3 and §2.3.4), for instance by directly engaging them (see Chapters 3 and 5). I thus highlight three research questions that constitute core challenges related to CRQ based on the latest human-centred literature on transparency: the Design Challenge (RQ1) [251, 379], the Guideline Challenge (RQ2) [35, 303, 304], and the Assessment Challenge (RQ3) [152, 197]. Although not exhaustive, I argue this list offers a starting point to improve transparency for non-expert users for they illustrate several key aspects of the principles-to-practice gap.

1.4 Three core challenges to building transparent algorithmic systems

I now describe the three sub-questions i.e. the core challenges to building transparent algorithmic systems that I investigate in this thesis in relation to CRQ.

1.4.1 RQ1: The Design Challenge

The Design Challenge concerns designers: it is often unclear for them *how* to design transparency in practice [304]. This challenge has been clearly identified in the literature [35, 379], and constitutes a key issue to address in relation to CRQ. For example, many technology startups are trying to move fast and disrupt their markets with innovative systems, but might not have the resources nor a specific focus on what transparency actually means for the users of their systems [251]. Several bottlenecks have been stressed, for instance, transparency is a fuzzy concept and thus needs clarifying [15]. Following recent Human-Centred Explainable Artificial Intelligence (HCXAI) research [109, 111], I argue that better describing the information needs of algorithmic system users is an essential step towards solving this Design Challenge.

1.4.2 RQ2: The Guideline Challenge

Moreover, new laws [343, 75, 258, 141] and guidelines [171, 364, 94] aiming to promote transparency are emerging, as well as more human-centred research in the field [108, 112, 372]. For example, the IEEE 7001 ‘Standard for Transparency of Autonomous Systems’ lists three reasons why transparency is important:

- Modern autonomous systems are designed to work with or alongside humans who need to be able to understand what the systems are doing and why (...) Transparency is important in adjusting expectations and, hence, building confidence.
- Autonomous systems can sometimes fail (...) Without transparency, finding out what went wrong and why is extremely difficult and may, in some cases, be impossible. Equally, finding out how and why a system made a correct decision is important for the processes of verification and validation.
- Without transparency, accountability and the attribution of responsibility can be difficult. Public confidence in technology requires both transparency and accountability. [171]

I reflect on the standards process making in more detail in §1.7. Likewise, the Online Safety Act that passed in the UK in 2023 imposes on providers that services be “*designed* and operated in such a way that transparency and accountability are provided in relation to those services.” (art. 1(3)(b)(iii)) [141] (Emphasis added). The role of design and designers in relation to system transparency compliance has thus been confirmed in recent laws and regulations [352]. However, mere calls for transparency have been shown to be ineffective [381, 121]. I thus argue transparency guidelines and design recommendations must align with designers’ understanding of transparency and users’ information needs to become

impactful (RQ2: Guideline Challenge). This challenge, although mentioned by others [303], has not been explored systematically yet. I argue that better outlining designers’ approaches and the obstacles they face when designing transparent systems is therefore a key starting point to overcome this challenge [213] and could constitute an important step forward towards answers CRQ. Toward this, I also argue the new transparency standards require more academic attention in relation to design practices [304].

1.4.3 RQ3: The Assessment Challenge

Lastly, despite its potential benefits, transparency is neither a panacea nor an end objective in itself [13, 79, 339], and can even lead to other risks, e.g. in terms of privacy [5, 29, 211], hence my focus on meaningful transparency in CRQ. To help mitigate some of these risks (though not all of them) and leverage the advantages of algorithmic systems, I argue transparency must become *meaningful* to users. Cobbe et al. [69] and Norval et al. [247] define meaningful information as *contextually appropriate*: it must be relevant, accurate, proportionate, and comprehensible to a given stakeholder. However, we still lack a methodology to assess to what extent a system is (meaningfully) transparent for users in practice (RQ3: Assessment Challenge). This challenge has been raised for example by Haque et al. [152] in a recent literature review, though I argue it deserves more emphasis as assessment methods could help provide practical tools to bridge the principles-to-practice gap. For instance, new laws call for transparency but often do not actually define how this is satisfied [352]. While there is unlikely a single solution to the Assessment Challenge, I argue insights from users and designers can help identify key considerations when attempting to address this challenge.

1.5 Thesis outline and main contributions

In this thesis, I provide empirical, critical, and theoretical insights that can help towards answering CRQ and addressing the challenges listed in Table 1.1. By directly engaging with various users and designers, I contribute towards bridging the gap between theory and design practice, so that algorithmic systems better cater for users’ information needs in real-world settings. This is key to make sure systems can be reviewed [68], challenged [265, 276], and used safely by users [110]. I acknowledge focusing on non-expert users—and on designers—can only be part of the solution to improve system designs for users, as other stakeholders are also involved, e.g. regulators and policy-makers. However, I argue this approach is needed as part of a joint effort from various perspectives (including from academia) to provide more practical support to designers and critical insights in this space. Indeed, there has been many efforts in studying transparency with a technical

approach [202], but there is still a gap in understanding transparency requirements from the perspective of the user.

Table 1.1: Key challenges and literature gaps in human-centred transparency research

Challenges	Literature Gaps	Thesis Chapters
Design Challenge (RQ1): It is often unclear for designers how to build transparent algorithmic systems	i) the need to better define non-expert users and their specific information needs in relation to algorithmic systems	Chapters 3 and 5
Guideline Challenge (RQ2): Recent guidelines and design recommendations must align with designers' approaches and users' needs	ii) the need to explore the gap between new transparency standards and design practices	Chapter 4
	iii) the need to better operationalise design recommendations from recent human-centred explainable AI (HCXAI) publications in real-world settings	Chapter 5
Assessment Challenge (RQ3): Lack of methods to assess (meaningful) transparency in practice	iv) the lack of methods to evaluate (meaningful) transparency	Chapters 3 and 6

To explore these three challenges, I delve into four related literature gaps, as summarised in Table §1.1. I describe them in more detail in Chapter 2. Note these literature gaps are directly connected to the Design, Guideline, and Assessment Challenges. For example, addressing the first two literature gaps (i.e. i) the need to better define non-expert users and their specific information needs in relation to algorithmic systems; and ii) the need to explore the gap between new transparency standards and design practices) contributes towards tackling all three challenges. Although non-exhaustive, this list of challenges and literature gaps illustrates the need for more research towards making transparency meaningful for users.

To explore the Design Challenge (RQ1), I start by focusing on better defining the users' information needs of under-studied users, e. g. parents in the UK in Chapter 3, and gynaecologists in Chapter 5. To help solve the Guideline Challenge (RQ2), I then investigate designers' practices and understanding of system transparency in relation to new transparency standards in Chapter 4. In Chapter 5, I also aim to help bridge the gap between theory and practice by testing and expanding design recommendations from recent academic research with users as co-designers. Lastly, I open new avenues to research the Assessment Challenge (RQ3) by benchmarking meaningful transparency in Chapter 3,

and experimenting with a critical method to assess a system’s transparency dimensions from the perspective of users in Chapter 6. As explained in Chapter 2, transparency is inherently context-specific. Therefore, while the insights may not be applicable across all socio-technical contexts, the objective is to provide relevant examples of how to address these challenges in specific, real-life settings.

Thus, I aim to contribute towards better supporting designers willing to build more meaningfully transparent systems for users in practice. My main contributions are both empirical (Chapters 3 and 5), critical (Chapters 4 and 6), and theoretical (Chapter 7) and include:

- Providing qualitative insights into the information needs of under-studied user groups, and what constitutes meaningful transparency to them (Chapters 3 and 5);
- Describing some key challenges and opportunities to improve meaningful transparency according to such groups and designers of algorithmic systems (Chapters 3 and 4);
- Understanding designers’ approaches to making transparent systems, and to what extent these align with international standard recommendations (Chapter 4);
- Testing recommendations from the latest HCXAI literature, with a prospective, real-life AI application in a relatively less-studied clinical domain (Chapter 5);
- Outlining research and design recommendations towards meaningful algorithmic system transparency for users (Chapters 4 and 5);
- Applying contextual integrity [246] as a framework to benchmark meaningful transparency in practice (Chapter 3);
- Introducing and experimenting with a critical method from art history to better account for the socio-technical context of meaningful transparency when assessing interfaces (Chapter 6).
- Reconceptualising algorithmic system meaningful transparency for non-expert users based on this thesis’ findings and contributions (Chapter 7).

I reached these contributions with real-life case studies (Chapters 3, 5, and 6) and interviews with professional designers (Chapter 4). In doing so, I hope to recenter transparency research on non-expert users’ needs and the practical challenges and opportunities faced by designers in charge of algorithmic systems. I now define the key terms, before outlining my conceptualisation of meaningful transparency and positionality.

1.6 Defining algorithmic system transparency, non-expert users, and design

This PhD thesis builds upon three core concepts: algorithmic system transparency, non-expert users, and design. I briefly define these and some key related terms (see words in bold below). All keywords and definitions are listed in the Glossary. I use accessible language throughout the thesis for a broad audience to benefit from my work, as exemplified in the Glossary.

Systems are “a set of interacting data, algorithm(s), and/or model(s) to form a technical workflow or product,” [204, 203]. I focus here on user-facing systems that include an interface and **Algorithms**, i.e. a computational method, formula, or procedure [203], for example a facial recognition algorithm in an automated border control gate that triggers an identity verification model [204]. Following Van de Poel, I define **Algorithmic systems** in this thesis as socio-technical systems “that are composed of technical artifacts, human agents, and institutions [and]—in addition—contain artificial agents and certain technical norms that regulate interactions between artificial agents and other elements of the system” [346]. Algorithms have varying levels of complexity and are employed in various types of **Artificial intelligence (AI)** [295, 203, 186], **Automated decision-making (ADM)** [76, 204] and **Machine Learning (ML)** [204, 203, 295]. Such systems can be applied in a wide range of domains, from education to **Clinical AI** [317], e.g. **Computer-Aided Diagnostic (CAD) systems** [101, 104]. AI is an overloaded, fuzzy concept [257, 203, 186]. It can be defined as “the designing and building of intelligent agents that receive percepts from the environment and take actions that affect that environment” [295], however what AI means can differ according to the context [203]. For example, several distinguish narrow AI, general AI, and superintelligence: “Applied AI in industry often refers to (1) *narrow AI*, the use of software to study or accomplish specific problem solving or reasoning tasks, in contrast to (2) *general AI*, which is able to reason, plan, and solve problems autonomously for cross-domain tasks beyond its original design, or to (3) *superintelligence*, a system that can outperform humans in every field” [203, 186]. To avoid ambiguities, I focus on algorithmic systems.

Transparency is also difficult to define, as it refers to a variety of related conceptual targets [15]. An oft-cited term [182], it includes different types of transparency [359], with different definitions according to the disciplines [35, 191, 106]. Even within computer sciences, there is no consensus as to what transparency means: it is often used interchangeably in the literature [68, 182, 18] with other related concepts such as **Explainable AI (XAI)** [112, 111] and **Interpretability** [107, 383]. For example, Erasmus et al. diagnose “the confusion about explaining neural networks within the machine learning literature as an equivocation on ‘explainability,’ ‘understandability’ and ‘interpretability’ ” [116]. On

the other hand, many complex systems can be opaque –often called ‘**Black-boxes**’ [317]– inasmuch as they “hide their internal logic to the user” [149]. In this thesis, I take a holistic, human-centred approach to help inform those willing to improve algorithmic system transparency and define it using IEEE 7001 ‘**Standard** for Transparency of Autonomous Systems’ (IEEE 7001) [171, 9] and Cobbe et al. [68] as a starting point:

Algorithmic system transparency: a transfer of information from an algorithmic system or its designers to a stakeholder that is truthful; contains information relevant to the causes of some action, decision, or behaviour; and is presented at a level of abstraction and in a form meaningful to the stakeholder. Transparency should be mindful of the stakeholders’ likely perception and comprehension, and should avoid disclosing information in a manner that, while technically true, is framed in a way that leads to misapprehension (source: [171] as available in summary in [364])

Meaningful transparency: a transfer of information to a specific stakeholder group that is *contextually appropriate*, i.e relevant, accurate, proportionate, and comprehensible to them (source: [68, 247]).

I use the IEEE 7001 definition for consistency in the thesis and because this standard represents my primary source to explore designers’ understanding of how to design transparent systems for users (see Chapter 4). I also focus on Cobbe’s definition of meaningful transparency to build upon previous research, aiming to improve transparency in practice. Indeed, transparency can help promote several values of **Responsible AI** [18, 122], such as human control or **Accountability**. The latter can be defined as “an institutional arrangement whereby an actor provides accounts to a forum, who deliberates on those accounts and may impose consequences” [70, 43, 42]. Without meaningful transparency, it would not be possible to understand, critique, and scrutinise algorithmic systems [68, 360].

For transparency to become meaningful though, it is important to recognise the key human stakeholders involved, primarily designers and users. This approach is shared by the latest **Human-centred Explainable AI (HCXAI)** [112, 107] publications, a recent branch of XAI [372] that “argues that explainability of AI is more than just “opening” the black-box— who opens it matters just as much, if not more, as the ways of opening it” [112]. For consistency, I also define as a starting point designers and non-expert users based on the same international standard IEEE 7001 [171]:

Designers: designers, developers, builders, maintainers, and operators, as well as decision-makers and procurers in organisations using and deploying algorithmic systems (collectively, “designers”) of algorithmic systems who either wish to or are required to engineer systems that have a certain degree of transparency (source: [171])

Non-expert users (users): people who have only a brief interaction or who interact every day with an autonomous system (source: [171]). I extend this to individuals directly or indirectly impacted by such interactions. By non-expert users, I thus mean non-expert users in AI, e.g. clinicians are medical experts, but (often) non-expert users in relation to algorithmic systems.

Lastly, I broadly define the **Design** of an algorithmic system in this work as the end-to-end process and final system resulting from conceiving, training, testing, and maintaining of an algorithmic model, as well as the associated software, interface and potential hardware (e.g. ‘smart’ toaster). In recent years, system design usually follows an iterative and agile approach. In this thesis, I delve in particular into the interface design of algorithmic systems as indicative of the system’s transparency dimensions (see Chapters 3, 6), and explore **Co-design** [299] as a method to ground, test, and expand transparency recommendations in real-world settings (see Chapter 5). For clarity in the usage of a common vocabulary, all the concepts in bold are defined in the Glossary. I acknowledge many of these definitions are limited and discuss them further in the following chapters. For instance, I argue that we need a more detailed and nuanced definition of non-expert users when considering algorithmic system transparency, as I discuss in the next chapter.

1.7 Reconceptualising meaningful transparency

In order to explore CRQ, I also aim to reconceptualise meaningful transparency for non-expert users, as initially defined by Cobbe et al. [68], with several theories and approaches (see §7.2.1). First, I use Nissebaum’s critical integrity to better define it in context, i.e. here of digital parenting in the UK (Chapter 3). I argue in Chapter 3 that it is one possible framework to benchmark meaningful transparency for non-expert users. Indeed, this theory has not been applied to conceptualise transparency so far, thus I suggest it as a way of moving forward in exploring the Assessment Challenge (RQ3).

To explore whether algorithmic system transparency can be quantified (which is also part of RQ3), I then explore IEEE 7001 with designers [171], and question whether the quantitative approach offered in the standard is likely to be implemented in real-life settings (Chapter 4). It is noteworthy that the standard’s making process is not explained nor the literature used to build its recommendations explicitly cited in IEEE 7001. I argue that this is an important limitation of IEEE 7001 when promoting transparency in system design, as it is unclear what data and/or theory its recommendations are stemming from. I therefore do not endorse this standard nor its methodology in this thesis, but rather use it as a starting point to define meaningful transparency (Chapter 2) and a prompt to explore CRQ with designers and users of algorithmic systems (Chapter 4). As a way to mitigate any possible risk of misalignment between such published guidelines and users’

needs in real-world settings (RQ2), I also provide methods and illustrate with real-life case studies how to directly engage with non-expert users (Chapters 3 and 5). However, IEEE 7001 remains one of the first international standards published in the field and, as such, I argue it is important to explore its definitions and transparency recommendations further.

To help bridge the gap between meaningful transparency and design practice, I then combine co-design methods, feminist HCI theory, and recent HCXAI design recommendations (Chapter 5). To my knowledge this is the first attempt to apply feminist HCI, as introduced by Bardzell [26, 27], to meaningful transparency in practice (see Chapter 7). I argue this is a productive body of theories to apply in the context of gynaecology as feminist considerations are not explicitly reflected in Cobbe’s definition of meaningful transparency and yet structural gender inequalities have been evidenced in clinical research and practice [55, 54, 276].

Lastly, I further explore a socio-technical system approach to algorithmic system transparency – as defined in §2.2.2– by introducing and experimenting with an interdisciplinary and critical method from art history to analyse the transparency dimension of algorithmic systems’ user interfaces (Chapter 6). To my knowledge art history as a discipline have not been used in transparency research nor HCI so far. I argue object-based analysis is particularly helpful in analysing the transparency dimensions of algorithmic system interfaces in contexts where users are not directly approachable, either for ethical or logistics reasons, as this methodology does not directly involve human participants. This is a major argument for using it in the context of our case study, i.e. the French national university admissions system, as explained in Chapter 5. Moreover, by analysing cultural and visual elements that might impact the public reception of a user interface and its transparency dimensions, I argue this art history methodology enables to account for a broader socio-technical context surrounding system design. As others before me, I argue a wider understanding of algorithmic systems as socio-technical systems is needed to bring transparency theory closer to design implementation (see §2.2.2). All these theories and methodologies fully align with my empirical, human-centred and impact-driven perspective and are directly linked to my positionality, as described below.

1.8 Positionality

This thesis is driven by my focus on leveraging Computer Science to increase my social impact as a left-/green-wing, queer, white European woman and feminist. For example, my focus on gynaecology in Chapter 5 is directly connected to my lived experience as a woman and gynaecology patient, whilst the APB and Parcoursup case study in Chapter 6 comes from my first-hand interaction with APB as a university applicant who went to (a state) school in France. Likewise, the interdisciplinary and co-design approaches I

use in these chapters are related to my commitment to feminist HCI methodologies, as introduced by Bardzell [26], as well as to my professional and academic training as an art historian (Chapter 6).

1.9 Ethical approval

This PhD project has been supervised by Prof Alan F. Blackwell and Dr Jatinder Singh. All the research conducted with human participants included in this thesis has been granted ethical approval from the ethics committee of the Department of Computer Science and Technology at the University of Cambridge between 2021 and 2023. Further ethical considerations will be discussed in each study.

Chapter 2

Literature review

As mentioned in Chapter 1, designing transparent systems remains an open challenge (CRQ). To explore this question, I start by justifying why transparency matters for users and unpack its definition (§2.1). I then describe the four literature gaps listed in Chapter 1 in more detail. In doing so, I also explain the difficulty of improving transparency in practice (§2.2) and highlight some main limitations to transparency (§2.3). More background knowledge specific to each study (e.g. co-design studies in HCI, or AI research for women’s health) will be introduced in the relevant chapter directly.

2.1 Defining and justifying transparency for users

Transparency relates to multiple concepts and fulfills many functions [121]. In this section, I justify why algorithmic system transparency matters for non-expert users and define it, mainly from a computer science perspective. I unpack key component parts of the IEEE 7001’s definition as a starting point: “a transfer of information from an [algorithmic] system or its designers to a stakeholder that is truthful; contains information relevant to the causes of some action, decision, or behavior; and is presented at a level of abstraction and in a form meaningful to the stakeholder” [171, 364]. For other approaches and definitions, e.g. stemming from philosophy, ethics, and politics, see [5, 23, 231, 381].

2.1.1 Opaque systems and associated risks for users

According to the IEEE 7001 definition above, transparency “contains information relevant to the causes of some action, decision, or behavior” of an algorithmic system [171, 364]. I argue this means that transparency matters to the user as a means to help them identify the potential risks involved with using a system, or diagnose what went wrong should they be facing some harm as a result. Indeed, one of the main arguments for improving algorithmic system transparency for users is to prevent and protect people from the potential risks

associated with opaque algorithmic systems, e.g. in terms of unfairness, discrimination [204] or privacy [291]. Such threats range in kind and severity and can affect individuals as well as communities and society at large [265]. For example, the Facebook files scandal in October 2021 epitomises how some systems can have a detrimental impact on a broad swath of users and not communicate about the risks they might carry for users [133]. Such issues motivate our focus on transparency for users.

Indeed, policymakers [343, 75] and academics from various fields [284, 113, 36] have worked towards reducing the risks algorithmic systems pose for users in recent years. For example, the UK passed in 2023 the Online Safety Act [141] to protect all UK-based users (and in particular children) by demanding more transparency from service providers, and similar debates about restricting access to smartphones and social media for under-aged users are currently ongoing in France [212]. Indeed, black box systems have become so ubiquitous that, in 2015, Pasquale [274] argued we had entered a “Black box society,” and called for a re-balance of this algorithmic power based on data processing on an unprecedented scale. Among academics who have endeavoured to address this challenge, Pedreschi et al. [275] introduced in 2019 a local-to-global framework for black box explanation towards providing meaningful explanations of Black box AI systems for automated decision making. Malgieri and Pasquale go further by advocating for “a presumption of unlawfulness for high-risk models, while the AI developers should have the burden of proof to justify why the AI is not illegitimate (and thus not unfair, not discriminatory, and not inaccurate)” [224]. However, the authors acknowledge this might not be easily administrable.

In practice, various types of opacity can currently be found in algorithmic systems [52, 84]. Cobbe et al. [68] for example list six types of opacity (Table 2.1), and reiterate that multiple form of opacity can be combined [68].

Table 2.1: Some types of opacity in algorithmic systems from [68]

Type of opacity	Cause
Intentional opacity	Details of processes concealed for commercial reasons [52, 84]
Illiterate opacity	Processes incomprehensible without technological literacy [52, 84]
Intrinsic opacity	Incompatibility between machine and human reasoning [52, 84]
Unwitting opacity	Unawareness of relevant of broader processes for accountability [68]
Strategic opacity	Process information deliberately presented in an inaccessible way [324]
Inadvertent opacity	Information unintentionally presented in an inaccessible way [324]

Algorithmic system opacity can also be associated with dark patterns, i.e. “user interfaces whose designers knowingly confuse users, make it difficult for users to express their actual preferences, or manipulate users into taking certain actions” [220]. This is particularly problematic since dark patterns have been proven to be widespread [352, 228]. However, dark patterns and opacity are not always connected, for instance when inadvertent

opacity occurs [325]. As a result, several actors have been working to regulate both opaque systems and dark patterns [352], and new guidelines are emerging to promote transparency.

2.1.2 Transparency goals at a high level

Moreover, IEEE 7001 describes transparency as “a transfer of information” [171, 364]. I argue this element reveals that transparency is about disclosure, and can function as a means of achieving other principles, in particular accountability. Transparency is indeed considered one of the core values of responsible and ethical AI: in 2019, Jobin et al. conducted a review of global ethical AI guidelines and found that transparency was the most cited principle in soft-law or non-legal norms, alongside “justice and fairness, non-maleficence, responsibility and privacy” [182]. Similarly, the OECD encourages transparency as a key value to support democracy and progress [150]. Moreover, transparency can be legally obliged, for instance in the UK’s Online Safety Act [141]. At a high level, transparency is often presented as beneficial:

- **Goal 1** “to minimise harm and improve AI” [182, 63] (§2.1.1)
- **Goal 2** for legal reasons [182, 123]
- **Goal 3** to foster trust [182, 166]
- **Goal 4** to promote accountability [69, 360, 93] (§2.1.4) and ethical AI values [182]

Thus transparency is often seen as a means to an end, not an end in itself [359, 339, 93, 121]. Turilli and Floridi call it “a pro-ethical condition for enabling or impairing other ethical practices or principles” [339]. I agree with the pro-ethical value of transparency and argue such property makes it a particularly interesting lever to improve system designs for users, as I aim to explore in this thesis. The goals listed above also relate to what Andrada et al. call *information transparency*: “the disclosure of information about an AI system, typically for the purpose of supporting judgements concerning the system’s fairness, trustworthiness, safety, efficacy, accountability, and compliance with regulatory and legislative frameworks” [15]. As the authors note, information transparency is important to support the formation of cognitions (e.g., judgements, decisions, beliefs) about a system’s suitability for use in a given context. Moreover, transparency can be used as a proxy to provide insights into characteristics that are difficult to measure, for instance robustness, safety, and fairness [359, 98]. Although largely accepted, this positive overview of transparency is challenged by several, as explained in §2.3. In any case, I argue transparency goals must be contextualised, as discussed below.

2.1.3 Transparency goals in context

Additionally, IEEE 7001 specifies that transparency must be “presented at a level of abstraction and in a form meaningful to the stakeholder” [171, 364]. I argue this element is key for non-expert users, as they encapsulate various audiences with different information needs. In practice, transparency thus needs to be tailored to a specific audience, can serve different goals, and take on different meanings according to the socio-technical context [359, 121]. By distinguishing different goals and stakeholders (see §2.2.3), Weller [359] lists eight transparency types, each requiring a different sort of explanation and measures of efficacy (Table 2.2).

Table 2.2: Weller’s [359] eight types of transparency

Type	Definition
Type 1	“For a developer, to understand how their system is working, aiming to debug or improve it: to see what is working well or badly, and get a sense for why.”
Type 2	“For a user, to provide a sense for what the system is doing and why, to enable prediction of what it might do in unforeseen circumstances and build a sense of trust in the technology.”
Type 3	“For society broadly to understand and become comfortable with the strengths and limitations of the system, overcoming a reasonable fear of the unknown.”
Type 4	“For a user to understand why one particular prediction or decision was reached, to allow a check that the system worked appropriately and to enable meaningful challenge (e.g. credit approval or criminal sentencing).”
Type 5	“To provide an expert (perhaps a regulator) the ability to audit a prediction or decision trail in detail, particularly if something goes wrong (e.g. a crash by an autonomous car). This may require storing key data streams and tracing through each logical step, and will facilitate assignment of accountability and legal liability.”
Type 6	“To facilitate monitoring and testing for safety standards.”
Type 7	“To make a user (the audience) feel comfortable with a prediction or decision so that they keep using the system. Beneficiary: deployer.”
Type 8	“To lead a user (the audience) into some action or behavior – e.g. Amazon might recommend a product, providing an explanation in order that you will then click through to make a purchase. Beneficiary: deployer.”

Weller’s transparency types 1, 4, 5, and 6 broadly coincide with Goals 1, 2 and 4 listed above, inasmuch as they aim to support safety, accountability as well as comply with

legal obligations. Likewise, types 2, 3, 7 broadly correspond to Goal 3. At first sight, transparency types 1-6 seem compatible with responsible AI values, however I argue the ethical benefits of types 7 and 8 are questionable, depending on the deployer’s aim (e.g. financial profit, widen access to public service, or state surveillance), and can arguably lead to risks for individuals and communities [5], when users’ and deployers’ interests diverge (see §2.3). As promoting accountability in algorithmic systems for non-expert users (Goal 4, type 4) has motivated this thesis, I now describe it in more detail below.

2.1.4 Transparency as a means to promote accountability

Indeed, IEEE 7001 stresses the need for transparency to be “truthful” [171, 364]. This aspect seems to imply that, for transparency to facilitate accountability, it needs to provide information that is backed by evidence and data and can be traced back to different logs and/or records. With this interpretation, I follow prior work that consider transparency as a means to facilitate accountability for various stakeholders of algorithmic systems [360, 249, 252]. Like Bovens [42] and Wieringa [360], Kroll [193] stresses that accountability is a relationship between several stakeholders, a key component also highlighted in the IEEE 7001 definition of transparency. Kroll explains that “one agent or entity is accountable to another for certain outcomes in certain contexts. Operationalizing that accountability relationship requires keeping records – accounts – of how systems operated and were created” [193]. In practice, accountability thus relies on accounts being created and communicated to a given stakeholder. Diakopoulos [93] stresses that the relation between the disclosing entity and the recipient shapes the effectiveness of algorithmic transparency in contributing to accountability. I agree with this emphasis on relationship, hence our focus in this thesis on making transparency *meaningful* for specific non-expert users in real-life contexts. Moreover, Floridi highlights that accountability is crucial to redress the current asymmetry between designers and users: “[the] situation is inherently unequal: a small fraction of humanity is currently engaged in the design and development of a set of technologies that are already transforming the everyday lives of just about everyone else” [123]. I build upon this approach and more recent work [99] on who is responsible for ethical AI by focusing specifically on the role of designers to effectively improve system transparency for non-expert users.

Indeed, the question of operationalising accountability is key: several concepts closely connected to transparency have been recently explored as means to do so in practice, for example traceability [194] and reviewability [68]. For Kroll, “traceability requires establishing not only how a system worked but how it was created and for what purpose, in a way that explains why a system has particular dynamics or behaviors. It connects records of how the system was constructed and what the system did mechanically to the broader goals of governance, in a way that highlights human understanding of that mechanical

operation and the decision processes underlying it” [194]. Operating at a different level of abstraction, reviewability has been introduced by Cobbe et al. as a framework for improving the accountability of ADM: “Reviewability (...) involves technical and organisational record-keeping and logging mechanisms that expose the contextually appropriate information needed to assess algorithmic systems, their context, and their outputs for legal compliance (...) or any other form of assessment relevant to various accountability relationships” [68]. I agree that uncovering “the why” [305] and organisational aspects involved in deploying an algorithmic system in practice is key for non-expert users. I build upon such body of work by arguing that, without *meaningful* transparency, it would not be possible to understand, critique, and scrutinise algorithmic systems. However, transparency and accountability are not synonymous [231]. For example, Ananny and Crawford [12] highlight the limitations of transparency, in particular for understanding and governing algorithmic systems. To clarify the challenges and gaps I have highlighted in the human-centred literature on transparency, I now summarise the relationship between transparency, interpretability, and explainability.

2.1.5 Defining transparency, explainability and interpretability

Finally, the IEEE 7001 mentions transparency should be “presented at level of abstraction and in a form meaningful to the stakeholder,” without specifying what “level of abstraction” and “meaningful form” might mean in practice. This matters because both can apply to transparency, as well as explainability and interpretability: all three conceptual targets can come in different forms and level of details, leading to possible ambiguities between these terms. I argue this IEEE 7001 phrasing is somewhat vague and thus constitutes a limitation to its definition of transparency. To further clarify what transparency might mean in practice, I now briefly disentangle the three connected yet different concepts of transparency, explainability, and interpretability. Indeed, there is no universally accepted definition of these concepts [116, 98, 236], so the relationship between them fluctuates across the publications: they are often used interchangeably in the literature [68, 116, 121, 18, 182], leading to possible ambiguity. Overall, explainable AI aims to provide human-understandable information for the system’s behaviour and processes [112]. Interpretability arguably goes further as it focuses on “provid[ing] explanations in understandable terms to a human” [383]. Reviewing all types of explainability and interpretability goes beyond the scope of this PhD thesis. Instead, I define explainability as a subset of transparency that focusses on explanations, as suggested in the international transparency standard IEEE 7001 [171, 364], and interpretability as explanations that are comprehensible to users. In other words, explainability includes various types of explanations, e.g. total or partial explanations, global or local explanations, whereas I define transparency as a broader concept: in principle transparency can encapsulate not only explanations, but any

type of information, such as general principles about data flows or system design.

This approach is oversimplified and the above definitions debatable. For example, some consider transparency part of explainability [111], while Diakopoulos [93] describes explanations (e.g. causal or case-based comparisons for a particular decision) and transparency (i.e. a description of the system behaviour and design intent) as distinct. Moreover, as Rudin highlights, models that are intrinsically interpretable do not need explanations [294]—though transparency is still required to determine whether they are indeed interpretable. Thus, I do not claim to reach consensus on these definitions. Rather, I use these as *workable* definitions and justify our focus on transparency as a broad umbrella term i) to navigate the literature providing relevant insights for algorithmic system transparency for non-expert users; ii) to situate our work in relation to this body of academic literature and align it with the definitions provided in our primary sources (e.g. IEEE 7001); and iii) to facilitate practical design improvement in algorithmic systems by using a single term, and thus make our work as accessible as possible to a wide audience. Since I consider XAI to be part of transparency research, this approach allows us to explore design recommendations for non-expert users from the recent HCXAI literature (Chapter 5). I now explain why it is difficult to improve algorithmic system transparency in practice, and further describe the four literature gaps I have outlined in Chapter 1.

2.2 The Difficulty of improving transparency in practice

Despite the recent changes in legislation, achieving algorithmic system transparency is recognised to be difficult by academics [68, 251] and designers [304], so many systems still lack in this respect. I now describe this ‘principles-to-practice’ gap and the four literature gaps I have highlighted in the human-centred research. For an overview of the challenges for making algorithmic systems more transparent, see [12, 87].

2.2.1 The Principles-to-practice gap

Recent HCI and Responsible AI (RAI) research has found a ‘principles-to-practices’ gap and called for more research on designers’ experiences to close it [221, 303, 240]. Indeed, mere calls to improve transparency have been shown to be inefficient [121, 381]. The role of designers in improving transparency in practice is thus key, hence the need to better understand their approaches and practices (Literature gap 2). Several academics [347, 35] have demonstrated this principles-to-practice gap by directly engaging with ML practitioners. I follow this body of work by interviewing designers (Chapter 4) when addressing the second literature gap, which I further describe in §2.2.4. This gap seems to

correspond to the limitations observed in other realms where transparency as a principle has been explored, for example in creating open source software. For example, when exploring open source fairness toolkits, Lee and Singh [202] show that several gaps between the toolkits’ capabilities and practitioner needs (e.g. a steep learning curve and other usability issues) hinder the adoption of such open source software in practice.

From a human-centred perspective, one key obstacle (among many others) that designers face is to understand what can be done in practical terms to improve algorithmic system transparency, e.g. in interface design [304]. This may be linked to the difficulty of providing examples of meaningfully transparent systems for non-expert users (see §2.2.5), as well as the Human-AI challenges stressed in Yang et al. [375], and may partly lead providers to prioritise illusions of transparency, rather than actual transparency design choices [352, 124]. Flyverbom [124] partly explains this by highlighting the complexity and “wider social processes and dynamics at work in transparency efforts.” While organisational practices and workflows are important to consider for transparency, in this thesis I mainly explore design artefacts and interfaces to focus on the relationship between the system and its designers and non-expert users.

To help circumvent the bottlenecks of improving system transparency, several academics have offered concepts such as ‘transparency by design’ [121, 227], or ‘explanation by design,’ i.e. “given a dataset of training decision records, how to develop a machine learning decision model together with its explanation” [275]. Felzmann et al. explain their nine transparency by design principles (proactivity, audience focus, risk disclosure, responsiveness etc) as a way to bridge the principles-to-practice gap between transparency and real-life design: “the complexity of transparency for ADM shows tension between transparency as a normative ideal and its translation to practical application” [121]. Similarly, Bhatt et al. [35] review ML explainability methods and conclude that the majority of the deployments are for ML engineers rather than end users; thus, “explainability” serves the needs of internal stakeholders, whereas the goal of transparency is to serve the needs of external stakeholders (i.e. the end users). In order to better serve external stakeholders, here non-expert users, I justify the socio-technical system approach I use in this thesis before describing the first literature gap I have outlined: the need to better define such users and their information needs.

2.2.2 Towards a socio-technical system approach

Historically, there has indeed been a focus on technical approaches to model explainability (XAI) for expert users [98], e.g. efforts to define explainable Machine Learning for professional developers [35]. Whilst these approaches can be valuable for such users, several academics have pointed out their limits in promoting safe and responsible algorithmic systems. Following Green [144] and Nouws et al [253], Dobbe and Wolters warn against the

risks of technical approaches to AI systems, in particular due to the potential associated harms: “the broader social and normative complexity of harms and the relation to design choices are naively narrowed down to a problem in the technical design of AI systems” [96]. Moreover, technical approaches have also been challenged for not answering the complex information needs of various users in real-life settings [35, 68]. Among others, Bhatt et al. attempt to address this issue by arguing that communicating the uncertainty associated with model predictions can complement explainability as a form of transparency [36]. And yet, transparency extends beyond a description of the model itself, and can comprise broader concerns such as organisation goals, or the intent of an algorithmic system [93]. This type of information has been increasingly highlighted in transparency research and, we argue, is necessary when exploring it in relation to non-expert users [107, 70].

This has led in particular to a broad acknowledgement of the central role of socio-technical contexts when defining or designing algorithmic system transparency. Among others, Zuiderwijk et al. [389] advocate for a socio-technical system perspective, by stressing the need for more domain-specific studies related to AI governance and the lack of empirical testing of AI systems in practice. Likewise, Dobbe and Wolters aim to provide empirical insights “to understand the extent to which existing design, use and governance practices for machine learning (ML) models are able to address the sociotechnical vulnerabilities of ML applications” [96]. I argue a socio-technical approach to algorithmic system transparency for non-expert users is key to bridge the principles-to-practice gap described above, as it can provide a more comprehensive understanding of both the risks associated with algorithmic systems and the complex information needs of users in real-life settings. I thus follow this approach in this thesis by exploring several real-life case studies involving different users and socio-technical contexts. My approach also builds upon the work of Cobbe et al. [68], who advocate for a holistic approach, beyond the one-to-one focus on a single designer-system-user type of interaction, to better review ADM involving Machine Learning. Similarly, Diakopoulos [93] stresses the impact of human decisions and actions on the ethical performance of a system, and how they are woven into and throughout ADM systems. I explore this aspect in more detail in Chapter 6, and argue a socio-technical system approach first requires a more nuanced conceptual understanding of non-expert users, as explained below.

2.2.3 Better defining non-expert users and their information needs (Literature gap 1)

As a reminder, I define non-expert users based on IEEE 7001 definition as: “[people] who have only a brief interaction or who interact every day with an autonomous system” [171]). I extend this to individuals directly or indirectly impacted by such interactions.

By non-expert users, I thus mean non-expert users in AI, e.g. clinicians are medical experts, but (often) non-expert users in relation to algorithmic systems. Burns et al.’s recent literature review shows that non-expert users’ specific characteristics and needs are not always sufficiently defined in the HCI literature [51]. Defining non-expert users is indeed not trivial. Moreover, the XAI literature focusing on non-expert users has been expanding in recent years [197], especially since 2018 [152]. However, the many technical approaches and tools [180, 187] aiming to promote transparency (e.g. datasheets [132], model cards [234]), rarely target non-expert users in practice [35, 304]. A user might want to interrogate a model, for example when being denied a mortgage or life-insurance policy [83]. It is also important to highlight that in some contexts, transparency for users might become problematic in case it encourages gaming (see a discussion of the Gale and Shapley [128] matching algorithm implemented in a university admission system in Chapter 6) or infringes on IP rights. For example, a pharmaceutical company might not want to provide any information to users about the ML-based systems they use for drug discovery for fear of losing its competitive edge in the market. In this thesis, I explore this gap by highlighting that different non-expert user groups require particular attention, and provide some as exemplars (Chapters 3, 5) to show how better understanding their transparency needs can help influence design. I also highlight that even within such groups, various users might have different transparency needs, for example in relation to their mental models, i.e. how they conceptualise algorithmic systems and make decisions in certain contexts (see Chapter 5).

2.2.3.1 HCXAI: Recentring transparency research on non-expert users

Indeed, one key factor to define transparency in practice is understanding the ‘who,’ i.e. the stakeholders receiving information from the system or its designers [364, 107]. It is accepted that transparency should be tailored to the needs of a specific stakeholder group [359, 15], as individuals’ information needs and levels of understanding of algorithmic systems can vary greatly [109]. HCXAI researchers specifically investigate the characteristics and information needs of such stakeholders [112, 111]. This human-centred approach constitutes a growing field of research, and has been used in various application domains, such as clinical AI [197, 355, 356]. I follow this approach in my user studies.

Moreover, several scholars call to go beyond the one-size-fits-all approach still common in XAI research [349, 19]. As algorithmic systems are increasingly involved in daily decisions across an increasing variety of domain areas [320], it is urgent to study in detail a wide range of non-expert users *in real-life contexts*. I aim to do so by exploring the transparency needs of users that have been under-studied in the HCXAI literature so far, e.g. parents of teenage social media users (Chapter 3), and gynaecologists (Chapter 5).

Nevertheless, studying non-expert users’ needs cannot solve on its own all the issues

related to algorithmic systems (see §2.3). For instance, the ‘many hands’ problem [70], whereby most systems are an ensemble of many systems provided by various deployers, likely requires a broader approach, e.g. by focusing on the relationship between the deployers responsible for a system as well as their relationship to users. This problem also highlights further questions about transparency regarding the appropriate level of detail users receive: for example, is it enough for them to know just which actors are involved? Or do they need or want to know who communicates with whom? I nonetheless argue that such questions are best explored in context by directly engaging with users, as I have aimed to do in Chapters 3 and 5. In doing so, I argue my approach fits within a joint academic effort towards practical improvement in algorithmic systems. Towards this, I also study transparency for user in interfaces (Chapters 3 and 6), as these constitute design artifacts that can reveal specific aspects of algorithmic system transparency, i.e. the relationship between the system or its designers and users.

2.2.3.2 Analysing algorithmic system transparency in user interfaces

As explained, transparency encapsulates various meanings and is a fuzzy concept. To help recentre transparency research on non-expert users and better align system designs with their information needs, I delve on analysing interfaces specifically in two chapters of this thesis. I treat interfaces as design artifacts that play a key role in the transfer of meaningful information about the system to non-expert users. While it is important to consider organisational practices and workflows as part of transparency research, my approach builds upon prior work that highlights that transparency represents a relationship between the system or its design and users, as discussed in §2.1.4.

Indeed, XAI visualisation constitutes a key research avenue to improve algorithmic systems for users, as highlighted in Haque et al’s 2023 [152] systematic literature review of XAI for end users. User interfaces represent a crucial site for transparency to take place for non-expert users. Unlike developers, deployers, and other experts, non-expert users might not be able to inspect the code behind a system, let alone understand it. Therefore interfaces constitute a crucial liminal space for them to interact with systems. Moreover, interfaces offer us a first step to look into systems, as transparency design limitations are often visible at the interface level. The focus on interface design has already been explored by academics studying the role of visualisation for model transparency [236]. However, as Rudin notes: “toolkits of ML algorithms offer little in the way of useful interfaces for interpretable ML methods” [294]. I go further by arguing that visualisation and interface design can support not only model explanation but also system transparency as a whole. This could be, for example, a dashboard for clinicians to interrogate a model predicting cancer diagnosis based on medical imagery –perhaps with a saliency map overlaid on a scan or MRI– but also to enquire more broadly why this model has

been deployed to help them with this particular diagnosis –perhaps by interacting with a chatbot providing layered information about the system’s development and deployment or seeing a timeline summarising the key stages of the AI lifecycle– [305]. Indeed, I argue that even an interpretable model can become opaque if the associated interface is not providing contextually appropriate information to non-expert users, and vice and versa. I now discuss the second and third literature gaps, which concern *how* transparent information can be transferred to users.

2.2.4 Exploring the gap between guidelines and design practices and operationalising HCXAI recommendations in context (Literature gaps 2 & 3)

While better defining users’ needs is an essential first step towards improving transparency, one key challenge for designers is to know *how* to make a system more transparent (Design Challenge (RQ1)). Towards this, recent guidelines and new standards have emerged, but these deserve further attention to make sure they align with both users’ needs and designers’ approaches to building transparent systems (Guideline Challenge (RQ2)). To explore these challenges, I outline in more detail the second and third literature gaps I have found in the human-centred transparency research: ii) exploring the gap between transparency guidelines and design practices; and iii) better operationalising design recommendations from HCXAI research in real-life settings.

2.2.4.1 Transparency tools, standards and designers’ approaches

As shown by Haque et al [152], XAI standardisation is a key area to explore further for users. Various design principles, tools, and frameworks have been suggested to support designers in making algorithmic systems more transparent [180, 187, 304]. Prominent mechanisms include datasheets for datasets [132], model cards [234] and fact sheets [318], which provide information about the models employed. Large technology companies such as Google and IBM also provide guidelines for AI development, including on explainability [270, 165]. Though different in scope and methodology, these attempts aim to facilitate specific types of transparency in algorithmic systems, and generally tend not to target non-expert users [304]. Several ‘transparency enhancing tools’ [180] have also been developed for users, e.g. Google dashboard [95], however these tools mainly focus on providing information about data collection, processing and/or disclosure, as opposed to system transparency overall. More guidance is thus needed [182], especially as designers still struggle to implement transparency and other legal requirements into systems [251, 252].

Alongside such tools, new standards have recently emerged in the field [63] and thus deserve further academic research. These new national [94] and international transparency

guidelines [177] could complement recent laws regulating algorithmic systems and data collection processes [75, 258, 141], such as GDPR [343, 304]. IEEE 7001 in particular offers transparency requirements, called Transparency Levels (TLs), in order to set “measurable, testable levels of transparency, so that autonomous systems can be objectively assessed, and levels of compliance determined” [171, 364]. However, as Jobin et al. [182] mention, the impact of IEEE standardisation initiatives on design practices needs to be explored further. Indeed, their recommendations need to align with designers’ understanding of transparency to become impactful. This literature gap motivates my investigation of designer’s approaches to promote transparency for users, alongside standards (Chapter 4).

2.2.4.2 Operationalising HCXAI recommendations

Similarly, the recent HCXAI literature offers several design recommendations for designers to make systems more transparent for users. These include for instance ‘personalising explanations,’ i.e. adapting to different kinds of people with different information needs [81, 64, 263, 215, 214, 197] –for example parents and children using social media (Chapter 3)– and ‘respecting professional autonomy,’ e.g. allowing clinicians use their judgment and decision-making in patient care without interference [354, 290, 356, 309]. These recommendations have been summarised in several literature reviews, such as Laato et al. 2022 [197], Wang et al. 2019 [355], and Wang et al. 2023 [357]. However, academics are calling for these to be better operationalised in real-world settings [112, 197, 110]. Ehsan et al. [110, 112] emphasises the need to operationalise the human-centred perspectives in XAI at the conceptual, methodological, and technical levels. For example, Haque et al. suggests more “research across multiple domains can also help portray domain-specific guidelines as well as generic guidelines for XAI development” [152]. I thus explore the challenge of implementing these HCXAI recommendation in design practice (Guideline Challenge (RQ2), Literature gap 3) with a real-life case study (Chapter 5).

2.2.5 Can we measure transparency? (Literature gap 4)

Despite some standards [171] providing recent attempts towards it, there is currently no consensus on how to assess the transparency dimensions of an algorithmic system for non-expert users (Assessment Challenge (RQ3)), let alone to assess meaningful transparency for users. Due to the principles-to-practices gap described above, it is indeed difficult to provide an example of a meaningfully transparent system for non-expert users in practice. Such example could help in choosing an evaluation method appropriate to assess other systems. Whilst HCXAI recommendations are working towards better aligning transparency goals with system designs and several transparency tools and interventions have been explored (§2.2.4), these still need operationalising and are not necessarily

generalisable since transparency is highly dependent on given users and socio-technical contexts. In other words, a system might be considered meaningfully transparent for one user group in one context, but less so in another context or for another user group, hence the need for more empirical studies in this space [305].

This constitutes an important bottleneck for building more transparent systems in practice, and it is unclear whether transparency can be quantified like other design properties such as usability, e.g. with the System Usability Scale (SUS) [47]. To my knowledge, no published research study has specifically addressed this Assessment Challenge (RQ3) so far (Literature gap 4). This research avenue has also been raised in Haque et al’s [152] recent XAI literature review. The authors recommend designing measurement scales and collecting user responses from different stakeholders. However, I argue their recommendation is not easily implementable given transparency is a fuzzy concept that encompasses a variety of related conceptual targets [15]. Beyond academic discourse, the IEEE 7001 standard precisely aims to set measurable levels of transparency [171, 364]. While discussing this quantifiable approach in Chapter 4, I opt in this thesis for a qualitative approach to explore this question from a human-centred perspective. I first introduce Contextual Integrity [245] as a framework to benchmark meaningful transparency for users in Chapter 3, and then experiment with an interdisciplinary, qualitative, and critical method to assess a system’s transparency dimensions in interfaces from the point of view of users in Chapter 6. Before diving into these users studies, I end this section by highlighting some important risks and trade-offs associated with transparency.

2.3 Addressing transparency risks and trade-offs

Although transparency is largely acknowledged as needed in algorithmic systems, several academics have outlined serious risks associated with it [5, 12, 87]. These span potential clashes with other responsible AI values, as well as limitations to transparency itself [359, 12, 87, 324]. I describe below some key trade-offs associated with algorithmic transparency, and how meaningful transparency can contribute towards addressing such limitations (§2.3.4). Note that I do not provide an exhaustive list; in particular, I do not discuss important colonial and ethical considerations related to transparency, for example its potential contribution to ‘ethical singularity,’ whereby “the evaluation of value systems assumes infinite value” [153].

2.3.1 Transparency vs Accuracy: a key trade-off to explore?

One of the main arguments for using black-boxes algorithms (e.g. deep neural networks, random forests) is their potential higher accuracy when trained with large datasets [217], e.g. for image recognition tasks. This focus on accuracy when developing algorithmic

systems is often considered to be in opposition with making systems more transparent or explainable. However, several academics have challenged the validity of this transparency vs accuracy trade-off [116, 294]. For example, some claim this is not a problem as such, since similar issues arise with human decision-making, i.e. these can sometimes be difficult to explain as well [217, 382]. Moreover, Rudin [294] demonstrates this trade-off is not always true, particularly when data are structured. They advocate for simpler, interpretable models such as logistic regression or decision lists, which can yield similar accuracy to more opaque systems in many contexts. Rudin adds: “Even for applications such as computer vision, where deep learning has major performance gains, and where interpretability is much more difficult to define, some forms of interpretability can be imbued directly into the models without losing accuracy” [294]. Lastly, Kroll [192] considers inscrutability associated with black-boxes a fallacy. They argue that “[the] mechanisms of a system’s operation can always be examined and explained, but a focus on machinery obscures the key issue of power dynamics” [192]. I agree with this approach, that also encourages to shift from post-hoc explanations to a more holistic understanding of transparency [305]. To sum up, I acknowledge this trade-off might apply in some instances, where, for example, more complex and opaque algorithms will outperform logistic regression. However, in such contexts, it falls to the developer to determine whether they should be opting for a more interpretable model, given their domain. As this question is out of scope in this thesis which focuses on non-expert users, I therefore do not explore this trade-off further, unlike the limitations and risks below.

2.3.2 Transparency vs Privacy: a complex relationship

Transparency has also been critiqued by several academics for its possible clash with other ethical values, in particular privacy [12, 87]. Indeed, Strathern [326] warns that transparency is not a neutral process, and strives to “avoid simply adding more visibility and more information” by stressing the current double bind of respecting diverse interests and increasing transparency. Similarly, Adams warns against the risks of state surveillance brought about by increased transparency in algorithmic systems [5]. However, I argue the relationship between privacy and transparency is complex. Although it is beyond the scope of this thesis to explore the relationship between transparency and privacy in detail, I acknowledge it is an important field with active studies [327, 142]. Instead, I highlight below some recent research adding some nuances to this potential trade-off. Seneviratne and Kangal [310] show that transparency can enhance privacy in some contexts, for example in healthcare or finance: the authors built “Privacy Enabling Transparent Systems (PETS) that makes transparency a key component in systems architectures” to circumvent over-broad permissions that may result in privacy violations [310]. By reviewing pre-2013 literature, Janic et al. similarly reveal that ‘transparency enhancing

tools’ (TETs) can give users “necessary information on how her data is being stored, exchanged, processed and used” and thus provide them with “better understanding of privacy policies, or provide insight in the third party tracking behavior” [180]. Again, it appears the relationship between privacy and transparency is context-specific. I explore this trade-off in relation to children’s online safety in Chapter 3. I now highlight two transparency risks that are context-agnostic.

2.3.3 Transparency fallacy & the transparency paradox

Two important limitations of transparency have been called ‘transparency fallacy’ and ‘the transparency paradox’. The first refers to the feasibility of designing transparency in practice. For example, when looking at European law, Edwards and Veale fear that “the search for a ‘right to explanation’ in the GDPR may be at best distracting, and at worse can nurture a new kind of ‘transparency fallacy’ ” [105]. They warn about the “danger of creating a meaningless transparency paradigm” for, they argue, such form of transparency may not be feasible and may not match users’ needs. Similarly, Berendt [29] analyses the AI Act [75] to provide reasons why transparency in itself does not solve the bias and discrimination risks brought about by algorithmic systems. They show that, “even if coupled with data quality, security, human oversight and documentation,” transparency is not sufficient to address all the risks posed by algorithmic systems, and thus should be considered only one among other web principles to thrive for [29]. Many scholars have raised similar points about transparency not being enough in and of itself, for instance recent research illustrating contexts, where transparency does not protect non-expert users from algorithmic bias [146]. We acknowledge this important limitation in each of our studies.

The transparency paradox, on the other hand, is what Stohl et al. [324] call the risk of opacity produced by high availability of information, approval to disseminate information, and accessibility of information to third parties. In other words, providing too much information to non-expert end-users might, paradoxically, prevent them from making meaning of such information. As a result, it is crucial to determine the right level of information for specific users (as mentioned in the IEEE 7001 definition of transparency [171]) in a particular context for it to become appropriate. This risk thus motivate my focus on *meaningful* transparency as a type of transparency that provides contextually appropriate information to non-expert users, as described below.

2.3.4 Towards meaningful transparency in context

Finally, I return to the IEEE 7001 definition of transparency, to further define what its mention of a “meaningful form” might mean in practice [171, 364]. In line with Cobbe et

al. [69] and Norval et al. [250], I move beyond this initial definition of system transparency to focus instead on ‘meaningful transparency.’ The authors define meaningful information as *contextually appropriate*, i.e. information that is relevant, accurate, proportionate, and intelligible to a given stakeholder. I argue that focusing on meaningful transparency is important to ensure transparency serves non-expert users needs in practice and because it can help mitigate the risk of transparency illusion described above. For example, I explore in Chapter 3 meaningful transparency as the transfer of information that should help parents i) understand and oversee the social media and online games that their children use and ii) protect their children online. However, I acknowledge meaningful transparency can only be part of the solution (both to help protect children online and to equip parents to better understand the online technologies that children use). To better define and assess meaningful transparency in context, I introduce Nissenbaum’s contextual integrity [246] as a framework to benchmark it. Initially created for privacy in the wake of surveillance technologies, I argue in Chapter 3 that contextual integrity can help characterise meaningful transparency. Nissenbaum [246] defines contextual integrity as the combination of norms of appropriateness and of distribution (or flow). The former norms relate to whether the content of the information provided is appropriate in a given context, while the latter dictate how such information should or should not be shared among various stakeholders. In the user study with parents in the UK for example, this means the information they receive about social media and online games must be i) appropriate for them to parent their teens and ii) follow the norms of transfer of information from one party (e.g. service providers, other parents, school etc) to them. I discuss this in more details in the next chapter.

2.4 Summary and main takeaways

Transparency is widely considered as needed, especially due to the risks posed by opaque algorithmic systems and their prevalence in our daily lives. It is therefore essential to improve transparency in such systems going forward, in particular for non-expert users who might not be able to inspect the code behind a system, let alone understand it. However, transparency is not an end in itself, is challenging to implement in algorithmic systems, and can entail some risks and trade-offs. Indeed, transparency is a complex concept that can fulfill various goals, and should be stakeholder-specific and contextually appropriate to become *meaningful*. By focusing on meaningful transparency for non-expert users of algorithmic systems, I build upon prior work on transparency as a means to facilitate accountability. Meaningful transparency is thus important to ensure transparency serves non-expert users needs in practice, and because it can help mitigate the risk of transparency illusion described above. Towards this, I have described the principles-to-practice gap

identified in the HCI and RAI research, and highlighted four gaps that require more attention in the transparency literature (see Table 1.1). Having unpacked the definition of transparency and some key arguments for improving it for non-expert users of algorithmic systems, I now explore these four literature gaps sequentially in the following chapters. In doing so, I aim to contribute towards bridging the principles-to-practice gap, and thus better align system designs with people’s information needs in real-life settings. Following the HCXAI call to focus on “the who” in explainability and transparency research, I start by directly engaging with non-expert users, here parents in the UK. I explore their transparency needs (Literature gap 1) in the context of their parenting responsibilities, as I argue such bottom-up, empirical approach is key to inform and ground transparency considerations in real-world environments.

Chapter 3

Understanding users' information needs: parents in the UK

This chapter is based on the following paper: Bianca G. S. Schor, Alan Blackwell, Jatinder Singh. 2024. Towards Meaningful Transparency for Digital Parenting. In *Journal of Responsible Computing* (under review). I conducted all of the data collection and analysis for this project, and was assisted in the writing and editing of the paper by my PhD supervisors Dr. Jatinder Singh and Prof. Alan Blackwell.

3.1 Introduction

In this chapter, I mainly address the first literature gap listed in Table 1.1: the need to better define non-expert users and their specific information needs in relation to algorithmic systems. Towards this, I explore the information needs of a specific user group, parents in the UK, in the wake of new online safety legislation. To the best of my knowledge, this is the first user study conducted with parents that focuses on algorithmic system transparency in the context of the recent change in legislation in the UK (i.e. the Online Safety Act). My aim is to provide guidance for designers, academics, and policy-makers on building and promoting systems that are meaningfully transparent for different types of non-expert users. Indeed, I argue that parents in the UK constitute an under-studied group of (secondary) non-expert users of online games and social media –the primary users being their teenage children– and thus deserve more attention. Again, I define non-expert users as “persons who have only a brief interaction or who interact every day with an autonomous system” [171, 364], including individuals directly or indirectly impacted by such interactions (see Glossary). I first justify my focus on parents in the UK.

3.1.1 Why focus on parents in the UK?

As their legal guardians, parents are responsible for children’s safety, both offline and online [167, 91]. Most children in the UK interact with online services before they turn 13 – i.e. the minimum age requirement to access most social media in the UK [342, 61, 91]. Children can benefit from using the Internet, e.g. to safely explore their identity or increase their social capital [91, 179, 341, 330, 33]. However, these services also entail many risks and potential *online harms*, i.e. an online behaviour that can hurt someone physically or emotionally, such as cyberbullying, grooming, blackmailing, or sending a person inappropriate content [261, 91, 264]. Though the impact of social media on teens’ well-being remains a controversial topic [267], children’s health can be damaged, leading in the most serious instances to depression, eating disorders, or suicide [264, 377, 348, 33].

I argue that for parents to effectively safeguard children online, the information they receive about such technologies must become *meaningful* to them. This parental duty relies on the expectation that parents will act in the child’s best interests, have sufficient e-literacy and information about the technology that children use (§3.3). However, previous research shows parents often face “black boxes” when managing children’s Internet use, i.e. they know relatively little about children’s online activities [139, 235]. Indeed, Blackwell et al. stress that the “practical obscurity” introduced by personal mobile use is one of the main challenges of digital parenting today [39]. Thus, while the online risks faced by children and trade-offs between parental oversight and children’s privacy have been studied extensively [264, 348, 377, 62, 34], parents’ information needs to address such risks, and their perception of such information warrants more attention.

Moreover, transparency is increasingly called for in the UK over the online technologies that children use. Indeed, the Online Safety Act passed in October 2023 with “a zero-tolerance approach to protecting children from online harm” [91, 103, 167, 141]. As this new law aims to “ensure social media platforms are more transparent about the risks and dangers posed to children,” and “provide parents and children with clear and accessible ways to report problems online,” the question of what constitutes *meaningful* information for parents in the UK becomes even more urgent to address.

3.1.2 Promoting meaningful transparency for digital parenting

To promote meaningful transparency for *digital parenting* —varied behaviours that contribute to parental monitoring and support in the postdigital age [235]—I first apply Nissebaum’s contextual integrity [246] to define meaningful transparency for parents (§3.2). I then profile parents’ attitudes towards information that might support them in overseeing technology that children use (§3.5). By highlighting parents’ perspectives, my aim is that those building systems (and those advocating for more responsible technology)

can make systems transparent in a manner that effectively supports digital parenting.

Towards this, I focus on parents' communities, as parents' experiences must not be considered in isolation [155, 283]. Indeed, parenting is likely influenced by socio-technical, cultural, and economic background [131, 378], and parents' communities deserve further academic attention [235]. I thus investigate the research questions: i) what types of information do parents find helpful about social media and online games to support digital parenting; ii) what are parents' attitudes towards such information; and iii) what are the appropriateness and distribution norms (see §3.2) between fellow parents that can help define meaningful transparency for digital parenting. Indeed, exploring such issues is not only crucial to help address the online harms that children are facing, but also urgent to better understand UK parents' role in safeguarding children online in the wake of the Online Safety Act. Building on Yardi's work [377], I focus on digital parenting as parents' understanding, adapting to, and monitoring of i) children's technology use, ii) the systems themselves, and iii) their impact on children over time (§3.3.3).

3.1.3 Chapter outline and contributions

To do so, I first define meaningful transparency for parents (§3.2), and how it can support digital parenting (§3.3). I then conducted semi-structured interviews with parents of 10-15 year olds in three communities in the UK, a country where concerns and debates about children's online safety are high (§3.4). In doing so, I attempt to account for a population comparatively less represented in the child-computer interaction and transparency literature [235] and use both oral and drawing activities to confirm parents' information needs. I focus on the 10-15 year old age range because it constitutes a key transitional period for children: in terms of schooling, access to technological devices and online services, and relationships with their families and peers (§3.3). Indeed, research shows children are particularly vulnerable when entering adolescence [264, 61], so meaningful transparency regarding social media and online games is essential for parents of 10 to 15-year-olds. I focus on social media and online games because they are the most common online services used by early teenagers [342, 261] and areas of concerns in the Online Safety Act [141]. I define social media as any technology children might use mainly to communicate online with friends (e.g. Tiktok, Instagram, Snapchat, Whatsapp) and online games as the online services that children use which contain an important gaming and playing dimension (e.g. Fortnite, Fifa, or Minecraft). In this way, I contribute towards:

- Applying contextual integrity [246] as a framework to define meaningful transparency for digital parenting
- Providing qualitative, empirical insights into parents' information needs regarding technology that children use

- Highlighting the sources and types of information that might be missing and helpful for digital parenting
- Describing some key challenges and opportunities to promote meaningful transparency for digital parenting, including design principles for the information flow between fellow parents.

3.2 Theory: Defining meaningful transparency for digital parenting

I now define in more detail meaningful transparency and contextual integrity [246] in the context of digital parenting.

3.2.1 Transparency for parents and its limitations

As discussed in §2.3.4, I define ‘meaningful transparency’ following Cobbe et al. and Norval et al.’s definition[69, 247] as the transfer of *contextually appropriate* information – i.e. relevant, accurate, proportionate, and intelligible – to parents. I apply this concept to analyse parents’ information needs regarding the technology that children use, to help them i) understand and oversee such systems and ii) protect their children online. There are many types of transparency relevant for parents, for instance transparency regarding a system, which could reveal how the system can impact young users; regarding a service provider and its business model; an algorithmic model and the data it is trained on; or regarding a specific child and their usage of such technology. However, transparency should not be conflated with control or surveillance mechanisms. Indeed, the question of what information is meaningful to support parents and for what purpose is especially important given the trade-offs between children’s privacy and online safety, which can result in tensions and conflicts between family members [138, 39]. While research shows the role of parental control software in children’s safety is complex [32, 137], I argue the question of meaningful transparency for parents should be explored beyond the context of restrictive parental apps.

The focus on making transparency meaningful is also motivated by the fact transparency is not an end in itself [3, 195, 255, 13, 324]. Although, it can lead to risks for children, e.g. intrusion into their privacy or surveillance [264, 136, 226], it is nonetheless largely recognised as helpful for supporting broader goals such as accountability [69, 360], as discussed in Chapter 2. I argue the risks children face, their age, and the multiple stakeholders involved in children’s online safety (e.g. online service providers, schools, parents, other users etc.) make the question of who is accountable for the online services children use particularly important to unpack; but without transparency over such technology, one cannot easily

challenge, critique, nor address how children might get impacted in the process [314]. I now apply Nissenbaum’s contextual integrity [246] to further describe meaningful transparency in the context of digital parenting.

3.2.2 Contextual integrity as a framework to define meaningful transparency for digital parenting

As introduced in §2.3.4, I argue Nissenbaum’s contextual integrity [246] can be used to support digital parenting by better defining meaningful transparency for parents. In doing so, I aim to open new avenues to research the Assessment Challenge (RQ3) in real-life settings and the fourth literature gap listed in Table 1.1, i.e. the lack of methods to assess (meaningful) transparency in practice. Nissenbaum defines contextual integrity as the combination of appropriateness and distribution norms: “norms of appropriateness dictate what information about persons is appropriate, or fitting, to reveal in a particular context (...) What matters is not only whether information is appropriate or inappropriate for a given context, but whether its distribution, or flow, respects contextual norms of information flow” [246]. I therefore use contextual integrity as a framework to conceptualise meaningful transparency and further operationalise it in practice: based on empirical insights, I identify norms of appropriateness and norms of information flow to help define what a meaningfully transparent systems might entail for parents in the UK today. Indeed, meaningful transparency must provide contextually appropriate information [69], I thus argue transparency becomes meaningful when it fulfills both norms of appropriateness and flow in a given context. In this chapter, I do not use contextual integrity to determine how data should flow appropriately between children, other online users, and service providers, though this could help raise issues related to privacy and surveillance. Instead, I focus on parents’ perspective within their communities to better understand their information needs for digital parenting: contextual integrity means here information received by parents about social media and online games must be i) appropriate for digital parenting (e.g. so parents can understand the technologies used by children) ii) follow the norms of transfer of information from one party (e.g. service providers, other parents, school etc) to parents. I discuss appropriateness norms in §3.5, and the distribution norms between fellow parents in §3.6. By outlining such norms for parents, I hope to help inform, guide, and promote parent-supportive design through improved transparency.

To my knowledge, contextual integrity has not been applied to transparency research so far. I argue it represents one among other possible frameworks that can help conceptualise and/or operationalise meaningful transparency. In this chapter, I show it is appropriate to outline norms to further define what meaningful transparency in practice in the context of digital parenting and parents communities in the UK. However, further research is needed

to evaluate this approach, for instance to assess how parents themselves consider such norms in relation to meaningful transparency. Similarly, more research is needed to use this framework in other contexts and to move closer to practical design implementation. I argue this chapter constitutes a first step towards this. I now discuss how meaningful transparency can help support digital parenting.

3.3 Background: how meaningful transparency can support digital parenting

This section provides background and outlines some key literature that demonstrates how meaningful transparency can support digital parenting, in particular to i) understand and oversee the technology that children use and ii) support, monitor and protect children online [69, 250]. While this section focuses on the risks for children using social media and online games (and therefore particular parental concerns around safety and harms), note that the benefits of such technologies have also been widely studied [91, 179, 62, 61]; these include, among others, developing a coherent identity and fostering communication and technical skills [330, 341, 33].

3.3.1 Meaningful transparency for whom? Parents’ role in children’s online safety & privacy

To explore meaningful transparency for parents, I first describe some key publications among the extensive literature on digital parenting [216, 65, 39] and parents’ role in children’s online safety and privacy. While I focus on parents’ perspective within their communities, recent papers focus on children themselves, e.g. to reflect their perspective on safety [135], or by involving them as co-designers of technologies that might support them [61, 230, 369]. Others have focused on learning technologies and schools’ influence on children’s interactions with technology and online safety [225, 46]. These interactions often raise complex trade-offs and tensions [39, 117], e.g. the key question of children’s privacy and parental control [74, 301, 135, 366, 82]. For example, Wisniewski et al. [22, 136, 296, 138] have investigated the communication and negotiation between parents and teens’ online safety: Akter et al. [8] recently tested a mobile application (“CO-oPS”) for collaborative parent-teen oversight, and showed it helped increase communication and transparency into their respective devices. However, CO-oPS revealed both parents and teens often believe parents lack the technological expertise to monitor smartphone usage themselves. Such research is helpful to inform future family-supportive designs, though the authors conclude “collaborative practices for managing online safety and privacy within families may be beneficial but also quite difficult to implement in practice” [8].

Among other issues, the conflicts triggered by children’s screen time remains frequent, and has been a key topic in the child-computer interaction literature. This, as well as personal, situational, and perceptual factors, can encourage the adoption of parental control software [323, 137]. Bertrandias et al. [32] have explored the impact of using such technology on parents’ well-being, and highlight how different features influence parents’ anticipated well-being. However, Ghosh et al. [137] show that more parental control is associated with more online risks. Thus, the role of such technology in children’s online safety is complex. Building on Ghosh et al.’s work [138], I explore parents’ information needs beyond parental control apps. Indeed, when discussing how families deal with children’s screen time, Hiniker, Radesky, and others [154, 155, 283] stress that parents do not bear the full responsibility of managing these challenging situations. They encourage us to move beyond “The Great Screen Time Debate,” and focus instead on the role of designers in creating equitable, child-centred digital environments on a population level.

Beyond parental control apps and screen times debates, parents’ communities require more emphasis in academic research: in a 2022 literature review of digital parenting, Modecki et al. stress: “parents often rely on a network of individuals (...) to monitor and support their adolescents online” [235]. By exploring the community influences on parents’ information needs, I build upon this systemic and community approach to promote meaningful transparency among fellow parents to support digital parenting. Given many recent digital parenting studies focus on the US [235] (Livingstone’s national UK survey was conducted 20 years ago [216]), I situate these issues with a comparatively less studied population, parents in the UK, where debates on children’s online safety are high.

3.3.2 Parents need meaningful information to address their legal duties and the risks children face online

Following the debates around the Online Safety Act [140, 141], parents’ information needs and children’s online safety have been a topical issue in the UK. The Act has attracted a lot of attention in the British media and public debate [147] and consists of a set of new measures aiming “to protect children and adults online.” It claims to “make social media companies more responsible for their users’ safety on their platforms (...) no matter where they are based, if they are accessible to UK users” [140]. As part of the UK’s efforts to improve children’s online safety, two national Ofcom reports from 2019-2021 further motivate and ground my research [342, 261]. Ofcom is the UK’s communications regulator, including the Internet, and is in charge among other things of online safety in the UK [259]. Ofcom’s regulatory work is largely recognised as important, e.g. in 2020, it became the regulator for online harms in the UK [260]. I aim to update previous studies [216] and to complement these broad quantitative studies with qualitative, in-depth insights.

As mentioned, social media and online games can benefit children: adolescent users report increased self-esteem, social support, social capital, and opportunities for self-disclosure [179, 62, 33, 341, 330]. However, Ofcom reports show whilst 75% of 10 to 15-year-olds declare using social media and over 80% online games in 2020 in the UK, the more time they spend on such services, the more likely they are to experience an online harm [342, 261]. Among the most common online services used by 10 to 15-year-olds today are social media (e.g. Tiktok, Instagram, and Snapchat) and online games (no specific game is cited in the reports), hence my focus on such services. Moreover, social media remain the most cited source of online harms. According to 12 to 15-year-olds in 2020, they are most often exposed to offensive language (39%), spam emails (31%) and unwelcome friend requests (29%), which they consider as harmful experiences.

Recent literature shows that online risks can lead to serious consequences on children's health and well-being, such as isolation, addiction, depression, and suicide [264, 348, 34, 61]; they can also become the victims of crimes and cybercrimes, ranging from privacy breaches to sexual predation [91, 226, 136, 365]. Academics have explored how to raise awareness of such risks among children and empower teenagers to address them effectively, e.g. Masaki et al. use nudge designs to assist teenagers in avoiding privacy and safety threats on social network services [226, 365, 61]. Several apps are also available to encourage healthy habits online, such as helping users to monitor their screen time [239]. As helpful as these designs can be, parents remain responsible for protecting their children by law (as their legal guardians), including online [167]. Given the frequency and potential negative impact of the risks and online harms cited above, it is therefore crucial to equip parents with the appropriate level and type of information to be able to do so.

3.3.3 Early teenagers are particularly vulnerable so parenting needs to adapt to oversee new technologies

Moreover, children are particularly vulnerable when entering adolescence [264, 61], so meaningful transparency is especially important for parents of 10 to 15-year-old social media and online game users to safeguard them online. In the UK, early teens transition from junior to senior school whilst the influence of peers, i.e. peer approval, comparison, and sense of belonging, becomes increasingly important for them. They spend less time with parents, distancing themselves psychologically, and foster new relationships outside their families. Teens start gaining more autonomy, which can cause parent-child conflicts [296, 136, 322, 102, 117]. These changes coincide with gaining access to mobile phones and using social media for the first time [342]. Moreover, adolescent cognitive development is associated with limited ability to self-regulate, making them vulnerable to technologies that can trigger addictive behaviours [61, 264, 34]. Such systems can affect children's

growth, (mental) health, and well-being, potentially leading to long-term negative impact [365, 179]. This motivates my focus on parents of 10 to 15-year-olds.

Lastly, parents’ struggles with digital parenting are not new, but the recent concentration of the digital economy has reinforced the information asymmetry they are facing [70]. Yardi calls “technoparenting” parenting teenagers’ technology use, and explore parents’ understanding of “what their teens are doing [on the Internet] and why” [377]. They show that, in 2010, US-based parents reported wanting more transparency over what their children were doing on their phones and online. At the time, most mobile phones in US households were not “smart,” so these two activities were distinct, unlike today [361]. Digital parenting is wider than technoparenting, as it encapsulates other dimensions such as parents’ use of technology in raising their children [297, 302, 235], however this demonstrates that parents in the 2010s already struggled with their unfamiliarity with technology and with understanding how to set rules for children, who were increasingly texting their friends. Thus, the rise in children’s mobile phone usage coupled with the dominance of few service providers such as Meta (formerly Facebook), Amazon, and Google over the last decade make increasingly important the question of meaningful transparency for parents regarding the online technology that children use.

3.4 Method

I now describe my method to explore parents’ information needs about technology that children use, their attitudes towards such information, and some key challenges and opportunities to improve transparency for digital parenting.

3.4.1 Parents and their communities

As discussed, parenting can be influenced by fellow parents and communities [130, 313, 235]. I thus focus on three communities in the UK by conducting semi-structured interviews over Zoom from April until July 2022 with parents of 10 to 15-year-olds. Despite the community focus, I opted for individual interviews to discuss such private, and sometimes painful, family matters. To profile parents’ attitudes towards such information, I selected parents whose children use social media and/or online games. Each participant comes from a different household and was recruited through a snowballing approach. Although I do not claim to represent the perspective of a whole community, this in-depth and local recruitment approach attempts to account for the influence of parents’ communities on their information needs.

The three communities were selected based on their geographic location, as individuals’ residence is correlated with social class in the UK (i.e. socio-economic, racial, and educational disparities). This can be estimated online via postcodes: one postcode is

shared by 15-20 households only in the UK [85]. They are respectively located in Shropshire (near Wales) [72], Cambridgeshire (East Anglia) [71], and London. The first participants for each communities were ex-colleagues and one ex-employer. Thus, while snowball recruitees from a same community know each other, e.g. many participants are neighbours, they are unlikely to know those from the other two communities.

3.4.2 Demographic information and ethical considerations

Table 3.1 summarises the main demographics of the participants: from 40 to 56 year old, interviewees have between 1-3 children. 61.5% identify as female and 100% live with their partners. Participants stated household income varies roughly between £60k and £500k/year, whilst their highest educational degree range from drama school diploma to PhD level. Self-declared technical literacy widely varies as well: some declare seeking occasional help when carrying daily online tasks, whereas one is a professional software developer. I am not interested in participants' exact level of income, technical ability, and education, but only relatively to other participants, hence the low, medium and high scale used in Table 3.1. This comparative ranking of income, education etc is also justified by ethical considerations, i.e. in order to protect participants' personal and sensitive information in a context where some other participants know them in real life. Compared to the Great British Class survey of 2011 [300], participants seem to belong to the elite, the established middle class, and the technical middle class. Moreover, several immigrant families were interviewed (e.g. Chinese, French), though race and ethnicity was not a selection criteria. Other ethical considerations regarding the personal and sensitive topics discussed with parents led us to opt for individual interviews despite our community approach, as described above.

3.4.3 Interview method and collaborative drawing activity

Since meaningful transparency is a new concept for many [250], I opted for an inductive, data-driven interview approach, whereby my questions related to the concept without mentioning the term explicitly. For example, interviewees were free to describe the types of information they found helpful about social media and/or video games, depending on their parenting experiences. I used semi-structure questions (see sample questions in Appendix A) as prompts to go beyond the discussion of parents' concerns (whilst addressing these as well) and information primarily for controlling children's online activities. Each interview lasted ~30-60 minutes and comprised three parts, all questions being optional. I first asked parents about their concerns and experiences regarding children using social media and online games to provide background for this study. We then discussed the sources and types of information they currently receive about such systems. I investigated what types

Table 3.1: Summary of participants' demographic information

Gender	Age	Nb children	Age/gender children	Technical literacy (self-declared)	Highest education degree	Yearly household income	Living with partner
Female	56	(4)	14 boy, 14 girl	+	+	+++	YES
Female	49	3	12 girl, 16 girl, 18 boy	+	++	+++	YES
Male	51	(3)	16 boy	+++	Not available	+++	YES
Female	52	(2)	13 girl	+	Not available	+++	YES
Female	40	3	10 girl, 7 boy, 3 girl	++	+++	++	YES
Female	41	3	13 girl, 11 boy, 8 girl	++	+++	++	YES
Female	41	2	10 boy	++	++	++	YES
Female	45	2	11 girl, 14 boy	+	++	+	YES
Male	48	2	14 boy, 17 girl	+++	+++	+	YES
Male	55	3	14 boy, 17 boy, 19 girl	+	+	+	YES
Male	51	2	14 boy, 16 girl	+++	++	++	YES
Male	46	2	12 boy, 13 boy	++	Not available	Not available	YES
Female	49	2	14 boy, 18 girl	++	++	Not available	YES

Legend: (2) total number of children in the family, including children from different parents
Income, technical literacy, highest education degree, yearly household income: + low ++ medium. +++ high

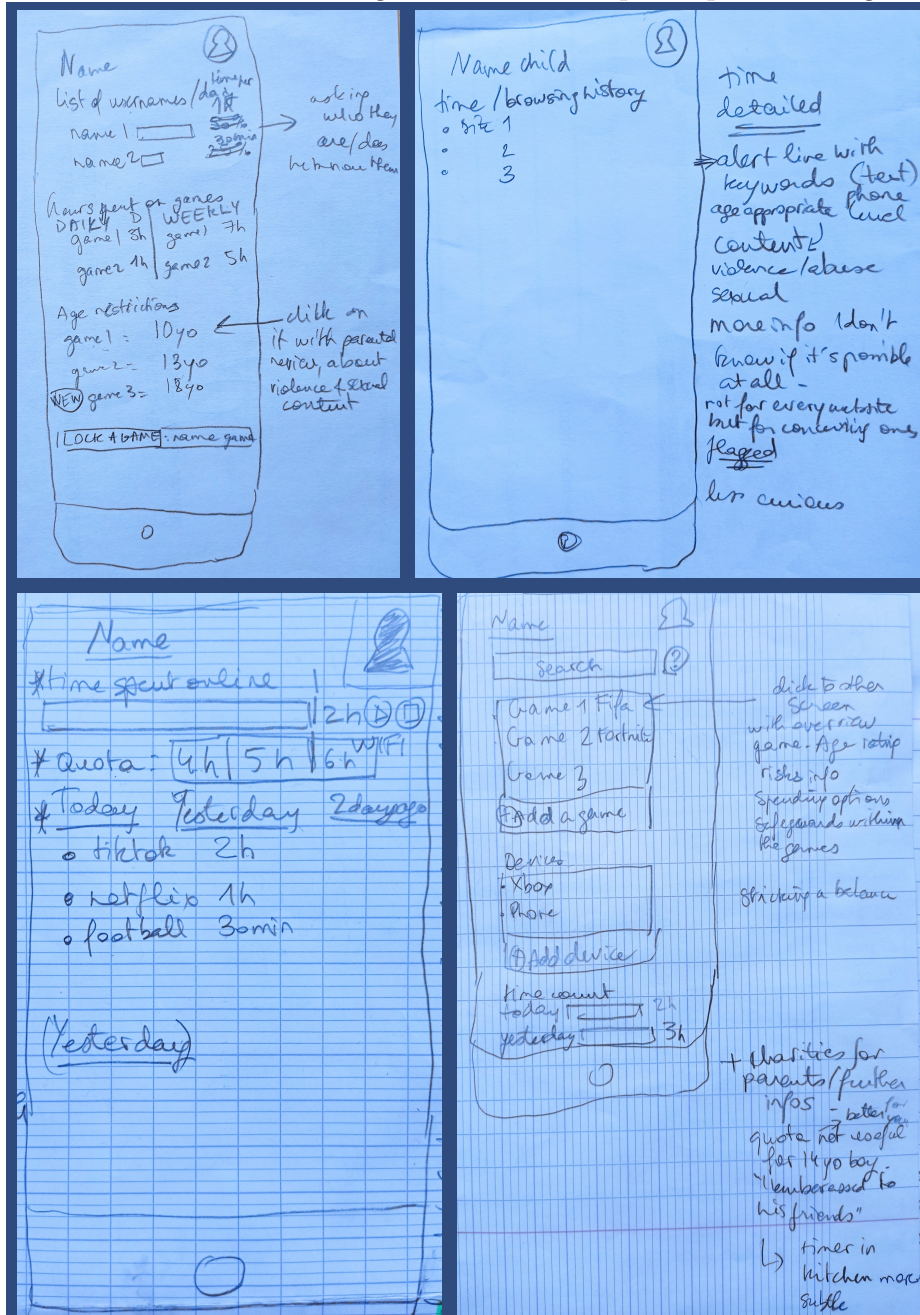
of information could be helpful for them in this context (in this way, to indicate what meaningful transparency might entail). As part of the interview, I experimented with a collaborative drawing activity as a second step to confirm, prioritise, and contextualise the types of information that parents had mentioned as helpful orally (Fig. 3.1). The drawings elicited further comments on how actionable such information would be for their digital parenting, which complemented and contrasted participants' initial statements.

The interviewer drew an exemplary user interface providing information for parents on children's social media and online games, and asked them to prioritise and confirm the types of information they would find helpful in practice on such an interface. I used a smartphone connected on Zoom to film the interviewer's hand drawing the interface with pen and paper, so participants could inform, guide, and correct it synchronously. Finally, participants could ask questions, were asked about their demographics, and to connect us with parents in their communities who might be happy to join the study.

3.4.4 Data analysis

I analysed the textual and visual data collected during the interviews using Braun and Clarke's thematic analysis, a 6-staged method for identifying and reporting themes in data [44]. I conducted all stages (i.e. familiarizing yourself with the data, generating initial codes, searching for themes, reviewing themes, defining and naming themes, producing the report) in two iterations. I searched for semantic themes in the interview transcripts,

Figure 3.1: Thumbnail of four drawings made live with participants during the interviews



my written notes, and the drawings, and then in the video recordings using an inductive, data-driven approach. The themes were then merged and validated, before repeating the process. I named the main themes based on interviewees' language and focused on themes raised by more than three participants (Table 3.2). Examples of codes include "parental review," "game overview," and "age rating" (see Fig. 3.1). By iteratively creating codes throughout the transcripts and focusing on meaningful transparency, parents' communities, and the appropriateness and distribution norms between fellow parents, I reached thematic saturation at n=13 interviews and 8 overarching themes. I now describe key findings from

my thematic analysis of the interviews and drawings with parents in the UK.

3.5 Key insights from online interviews with parents

Table 3.2 summarises the main themes raised by at least three participants. My main empirical contributions are summarised in rows 1-3 (Parents' perceived level of information and attitudes; Info helpful to parents; and Context & other stakeholders). I now describe these themes and four appropriateness norms I identify for digital parenting.

Table 3.2: Summary of the main themes raised by participants

Overarching category	Themes
Parents' perceived level of info & attitudes	We have no idea/don't know much (and we are aware of this); Our responsibility to know more/monitor, we need more info; It is not realistic/possible to control social media/online games; We are too old, it moves too fast, we have to catch up, we don't use it; We accept/give up; Not concerned maybe because not aware; The info is there but it requires to be active, it is hard to implement/too time-consuming; Info is overwhelming or hard to understand; Positive sides of tech: communication with friends, covid
Info helpful to parents	Time-related: time spent online and/or on a specific site/game, browsing history, time quota, option to block access; Content-related: type of content, age-appropriate content, level of violence, age rating, browsing history, alerts and option to block access; People-related: identity of online contacts; Money-related: in-app encouragements to spend money; Software or device specific: general information about software, time or software on a specific device; Parent-related: parental reviews or parenting advice
Context & other stakeholders	Schools: give online homework, resources and/or tablets, control them, raise awareness, send information, aware of issues, not involved enough; Other parents: talk together, not involved with them, parenting blogs/forums, solidarity; Companies: responsible for addiction; Public transport: commuting time spent on phones/games; Government: should step up with Internet policy; Culture: boys & girls, Asian & British
Devices & software used by children	Social media: TikTok, Snapchat, Instagram; Games: Fortnite, Fifa, Minecraft, Wordle, Road Blocks; Devices: phones, Xbox, computer, tablet, TV; Other: Youtube, Whatsapp, Netflix
Parents' concerns about children	Design-related: too much time, addictive, privacy, in-app transactions, waste of time, algorithm; People-related: grooming and interactions with unknown (adults), cyberbullying, lack of (real) human connections, inability to disconnect and/or switch back to reality; Content-related: violent, inappropriate, stupid, pornography; Short-term negative impact: aggressive/hyped up behaviour; Long-term negative impact: limiting/fake info or perception of life/the world, distraction from education; Health-related: stress/anxiety, bodily appearance, tense
Parents' coping mechanisms	More directive: ban some games/sites, parental control services, monitoring time/use and browser history, access or remove the device; Less directive: talk to child, trust them, partner monitors, give reminders, child asks for in-app purchases, talk to partner, have used the games/social media, want to delegate to the service providers
Family values & social interactions	Values: trust and open relationships, equip child with tools to address risks, books, sports and/or education are more important; Social interactions: we talk, we ask directly, bank account/email notification, elder teen influences/tells younger, generation conflict, deeply affects family life, adults giving the example
Children's needs & attitudes	Needs: freedom, privacy, interaction with friends; Attitudes: don't know the dangers & bad sides, override the control/limits, delete apps, decrease use or block/report people themselves, use peer pressure argument

3.5.1 Participants know little about the technology that children use but are keen to understand them better

Although many participants feel they do currently not comprehend social media and/or online games used by children, several seem keen to improve their understanding of it. This is encouraging for parent-supportive design through increased transparency. Most declare not having information about what is happening behind children’s screens (“to be honest, what I only know is that it’s a football game, I don’t even know the specific name” P5), and are aware of their limited understanding (“[I am] generally very aware of my ignorance on social media and video games” P10), but several explicitly say they need more information to better understand social media and online games: “it would be good to feel that we understood the games a little bit better, and the risks I suppose, and the safeguards as well that are in place, I would hope they are some” P2. Another parent even asks us for explanation on how social media works: “I think it’s very difficult on Tiktok to... sieve through, there’s so much on there, I don’t know... do you know much about Tiktok? Are you aware of it?” P6. However, this attitude is not unanimous: a few participants confess they have given up trying to understand the technologies children use. This should also be accounted for when designing parent-supportive tools to reach population-level change (§3.6.1), as called for by Radesky and Hiniker [283].

3.5.2 Missing information in participants’ current approach to digital parenting

One type of information seems to be missing in participants’ current approach to digital parenting: information about the systems’ design. Table 3.2 confirms previous research [216, 154] showing that parents’ attention lies primarily on children’s technology use (e.g. screen time, in-app purchases) and the potential negative impact of such interactions (e.g. grooming, anxiety). While several participants raise concerns about the design of social media and online games, and consider the companies as responsible for children’s addictive behaviour online, these concerns are not clearly connected to an explicitly stated information need. For example, few participants mention information about the algorithms underlying social media or online games as helpful for them to parent their children online. The drawing activity seems to confirm this discrepancy between participants concerns’ and the information they mention as helpful: none of the top-of-the-mind information they include here are about the systems’ design (Fig. 3.1). This gap is likely due to participants’ lack of understanding of and self-deprecating attitude towards such technology (see §3.5.1), however it also epitomises the need to better include users in the product lifecycle so as to reduce the information asymmetry between them and designers and make sure technologies are built and used to serve people [123]. One participant states “as parents we’re naïve

because we’ve never had this technology when growing up, so . . . it’s a bit of a minefield” P2. This attitude seems deeply-rooted and widespread among participants, regardless of their demographics, and almost inexorable for some: “[my son] would be on his phone, playing some game, which I completely don’t understand because I’m too old” P4.

This can be framed as an appropriateness norm of meaningful transparency for digital parenting: i) curiosity-supportive, i.e for parents to be able to ask specific questions about the design of algorithmic systems, and thus transparency to be improved by supporting interrogations of such, the information they receive should help reassure them on their ability to understand and question such information. Were levels of understanding greater, some might demand more information: “Actually, if I know more, I would like to make more criteria, but so far . . . I don’t have any other idea to do [so], I want them to know about how other people are playing in the safe . . . in the safe mode I would say” P5. Among the other types of information mentioned as helpful in the drawing activity, parental reviews about a specific online service could provide an opportunity to address this lack of understanding, with more tech-savvy parents sharing knowledge with others in their community (§3.5.4).

3.5.3 Several participants feel overwhelmed with the information they receive regarding their children

Providing parents simply with *more* information over technology children use is not necessarily helpful to them. Some already feel overwhelmed by the amount of information they currently receive about their children in general: “you get so much information all the time when you have children, so much information, you know, I probably get 50 emails a week from their schools (. . .) it’s completely overwhelming the amount of information you get all the time” P11. Indeed, many acknowledge not having digested nor used the information about their children they receive via email from schools. This echoes a general concern with regards to algorithmic systems, which Stohl et al. call the “transparency paradox:” “high levels of visibility of information decrease transparency and produce opacity” [324]. It thus further confirms the needs for transparency to be proportionate to parents’ information needs, and contextually appropriate [69].

This leads us to identify another appropriateness norm of meaningful transparency for digital parenting: ii) concision & precision, i.e. information needs to be pragmatic though not overwhelming in quantity. Indeed, several participants declare that information about their children’s interaction with social media and online games is available, e.g. receipts for children’s in-app purchases or reviews of specific online games, but can be too time-consuming and/or hard to retrieve or understand for them: “if I wanted to spend hours and hours, I could go through and monitor the different devices (...) I do have

access to information. I am overwhelmed by the monitoring that would need to happen because it's multiple devices, different places (...) it takes a hell of a lot of effort" P13. I now consider testimonies that can inform community-based design for digital parenting.

3.5.4 Communities provide most information on technology children use, seemingly not service providers

Most participants declare getting most and various types of information from their communities, who thus provide a fertile ground for parent-supporting design. These are mainly local, but also online communities of parents, e.g. on parenting forums. This information includes:

- General information about online safety and privacy, often from police officials, via the children's schools: "There's two ex-policemen or something that go around giving talks about online safety, and safety on social media and online games (...) I think that's been pretty good" P10. However, two participants declare not receiving any information from schools on such topics, and/or complained these are not involved enough. Besides, relying on the police as a source of information about children online safety and privacy has limits: the police may focus on the online harms associated with illegal behaviour and (cyber)crime, as opposed to other potential risks linked to the software design and the psychological impact of using technology, such as addictive behaviours or unwanted in-app transactions. Thus, the role of educational institutions in supporting successful digital parenting seems currently dependent on individual schools and deserves further investigation [235].
- Advice and information about a specific software or technology-related issue from other parents in their communities, their partners, or their own children: "[My son] knows more than I do" P11 (see also [117]). However, this can become challenging for parents: the same participant tells us "I felt he was manipulating me." Most research on digital parenting focuses on the communication within families [296, 377, 327], e.g. [190] calls "teen gurus" adolescents who have more technology knowledge than their parents and provide technical help. As Modecki [235] highlights, bidirectional influence (including from child to parent) and communities' impact on digital parenting still deserve more attention (see §3.6).
- Digital parenting guidelines and advice in parents' blogs and forums (§3.5.5). One participant cited commonsensemedia.org and askaboutgames.com in a post-interview email [345, 344].

Two additional appropriateness norms for meaningful transparency for parents appear here: iii) easily actionable, i.e. information should readily support digital parenting inter-

ventions; and iv) experience-based, i.e. information should relate to parents' experiences of overseeing technologies used by teens. On the other hand, no participant mentions receiving information from service providers, except those who directly access and monitor the devices and/or accounts their children use, e.g. by looking at the messages on children's phones. Some report realising their children were using online games in an unwanted manner, only when receiving bank statements or email receipts for unknown purchases: "we honestly had no idea he was even playing Fortnite, that's how scarily secretive it can be. We found out purely because there was a bill (...) Had we not found that bill, I promise you we would never have known" P6. This seems to indicate a lack of communication between service providers and parents. Although this is a legal obligation for service providers processing children's personal data, in practice several interviewees report their children did not ask for parental consent to create social media accounts. [167]. As one participant notes, tech companies target teens, not parents, despite parents being primarily responsible for them online (§3.6.1). This highlights a real opportunity for providers to do more for children's online safety, and meaningfully transparent designs for parents can help to do so.

3.5.5 Participants find the information they receive from fellow parents helpful

Among the information they receive, participants say the information coming from other parents is helpful to them. This attitude was not unanimous, but seems to encapsulate both online parenting communities and local parents. Participants who mention having received information from schools seem more ambivalent about how useful this has been. For example, one participant recalls finding the workshop they participated at a primary school helpful at the time, but when asked about the guidelines they had co-produced with children, declared not having them anymore. On the other hand, the same participant quickly found the parenting blog post they had used to create their household screen rules. This example illustrates the potential for online design to support digital parenting, as online information might potentially be easier for them to access and retrieve (see Table 3.3), in particular design based on the flow of information between fellow parents. One participant raised the idea of having a Whatsapp group with fellow parents in their community to discuss such topics: "I think maybe we need to strengthen... our communication with the other parents (...) if there was a way for us to really get some transparency on all of the different chatter around what our sons are doing when they are gaming together... and swap notes on it and experiences, that would be a good thing" P13. However, they had not done so themselves, as it would possibly be "extra effort for us (...) it's [about] making the time and the space to do it." After providing qualitative

insights into UK parents' information needs and identifying four appropriateness norms for meaningful transparency for digital parenting, I now discuss our results and six distribution norms for the information flow among fellow parents.

3.6 Discussion

I now situate the findings in relation to the role of parents, service providers, and communities in protecting children online. I then suggest community-based design opportunities to improve transparency for digital parenting by outlining some norms of flow among fellow parents in the UK, before describing our study limitations.

3.6.1 Parents vs service providers' responsibility towards children's online safety

In this study, I have expanded the research on parents' communities [235, 130] by directly engaging with them, and advocate for a community-centred approach to parent-supportive design (§3.5.4). Such participatory approach can help foster solidarity and support among fellow parents (Table 3.2, §3.5.5), so they can better face the challenges of digital parenting. Meaningful information should be available for those who want to be actively involved in safeguarding their children online. However, this can only be part of the answer (§3.5.1). Indeed, many factors can affect parents' ability to safeguard children online, e.g limited e-literacy ([125]), ethnic and socio-economic factors [378, 131], and historically parents have struggled with digital parenting [377, 216, 154]. Digital parenting should evolve alongside new technology (§3.3.3, [269]), but there needs to be a (legal and design) baseline for children whose parents are not going to use this information, so that no child ends up unprotected online. I thus argue that if providers engaged parents and their communities more, this would help improve i) transparency from the outset, ii) parents' e-literacy and understanding, and thus iii) their digital parenting overall.

Indeed, I show the information flow from online service providers to parents seems to be currently lacking (§3.5.4), which emphasises that the former can do more for children's online safety by making their services more meaningfully transparent for parents. This matches the standpoint adopted by the Online Safety Act in the UK [141], which aims to enforce a duty of care on service providers (§3.3.2). This raises the question of the incentives for tech companies to provide more information to parents, as it might expose them to further scrutiny. I argue that a growing incentive, for instance in several European countries like the UK and France [16, 141], comes from new legislation demanding more transparency from service providers and an increased focus on children's online safety in public debate and the media [147]. Tech giants are certainly more influential in providing

the information to safeguard children online than individual parents. Indeed, many problems associated with children technology uses are systemic. Several participants call out service providers for encouraging addictive behaviours (Table 3.2), and advocate for the government to reach better Internet policies: “I think the rule-makers and the governments are really many, many, many steps behind unfortunately (...) there is a strong case right now for better policing of what’s going on on the Internet” P8. Involving parents directly in policy discussions can hopefully promote laws that better benefit children.

3.6.2 Communities as a starting point for parent-supportive design: norms of flow among fellow parents

By profiling, beyond parental control apps [138, 22] and screen time debate [155], parents’ attitudes towards the information that might help them understand and oversee the technology that children use, I contribute to updating previous studies (§3.3.1) and complement quantitative surveys [216, 261, 342] with qualitative insights in the UK. I do not claim to exhaustively cover the topics of digital parenting and meaningful transparency for parents, but suspect our findings may extend beyond the British context.

I also apply Nissembaum’s contextual integrity framework (§3.2.2) to better define meaningful transparency for parents, by describing appropriateness norms for digital parenting (§3.5). To move forward, I argue communities can constitute a starting point to support digital parenting through design (§3.5.5). While participants consider fellow parents a source of helpful information (§3.5.5), the information they report receiving is scattered across various offline and online communities (§3.5.4), rarely comes from official sources, and is mostly in written or oral format. The flow of information among fellow parents could thus be improved by clarifying, centralising, and simplifying information on social media and online games they receive, as illustrated by P13’s comment (§3.5.5). Based on our findings, table 3.3 lists six distribution norms among fellow parents I have identified in terms of meaningful transparency for digital parenting in the UK. Though not exhaustive, this list deserves further research to be confirmed, and, I argue, can lead to principles for community-based design supporting digital parenting. For example, the information could be centralised through an online forum or channel similar to a community-based Whatsapp group, as one of our participants suggested (see §3.5.5).

3.6.3 Study limitations

3.6.3.1 Limitations related to our methodology.

Due to our choice of in-depth research methods, our study has been completed with an informative but not representative sample of parents in the UK, e.g. in terms of

Table 3.3: Non-exhaustive list of 6 distribution norms governing the information flow between fellow parents (in random order).

Distribution norm	Norm description in the context of meaningful transparency for digital parenting in the UK
Bi-directional clarity	information needs to be understandable both by parents providing and receiving it (Table 3.2, §3.5.1)
Timely & high frequency	information needs to be shared regularly and at the right time for parents to keep up to date (Fig. 3.1, §3.5.5)
Effortless & time-efficiency	information should easily and quickly flow between fellow parents (who are often busy) (§3.5.3)
Centralisation & directness	information should be centralised and directly retrievable by parents (Fig. 3.1, §3.5.3)
Parent-specific & culturally sensitive	the backgrounds of parents providing information should be clear to parents receiving it (Table 3.2)
Solidarity	information should be based on solidarity and promote support among fellow parents (Table 3.2, §3.5.5)

demographics. I acknowledge that some participants might have had a potential self-selection bias, e.g. were perhaps more interested in discussing digital parenting issues than others. Nonetheless, thematic saturation was reached at 13 interviews. I also acknowledge our sample does not represent all social classes in the UK, however this study was informed by two national quantitative surveys from Ofcom [342, 261]. I thus opted for a more restricted sample to explore these topics with a more in-depth, qualitative approach. Further research is thus needed on the information needs of other social classes, in particular working class parents, as I argue that all communities and families should have access to transparency that is meaningful for their digital parenting styles. Moreover, due to the small sample size, it was not possible to compare the finding across the three communities (the results would not be significant), though I was able extract information on parents' information sources, which primarily included their local communities. Furthermore, recruiting participants to discuss such private, and sometimes painful, family matters proved challenging, hence my decision to carry out individual interviews despite the community focus. I acknowledge participants' answers are influenced by their partners, e.g. their opinions and digital parenting styles. My findings should therefore be tested and expanded with other groups of parents. Though not the aim for this study, this method also meant I did not collect sufficient data to explore the influence of parents' cultural, ethnic, religious, sexual backgrounds and preferences nor single parenting, though these have been proven to influence parents' attitudes towards technology children use [313, 131]. More research is needed on such influences, for instance with a feminist approach [26, 27] to reconceptualise the role of different family members in digital parenting.

3.6.3.2 Limitations related to transparency.

Moreover, transparency in itself has limits [195, 324, 5, 13]. Our study confirms the “transparency paradox” [324], for providing too much information was described as overwhelming by parents, and can obfuscate the nature of technology children use.

Transparency, like in any context, becomes helpful only as a means to an end, i.e. here, when the information parents receive increases their ability to achieve successful digital parenting. Several factors can influence to what extent families can benefit from it, for instance whether parents are well-intentioned and virtually always behave in children’s best interests. In this study, I did not investigate the risk of transparency having dual uses, e.g. by increasing the risk of intimate partner surveillance, as it was beyond the scope of this study to investigate in detail the communication and potential power dynamic between parents, though I acknowledge further research is needed to explore such risk going forward. Although it goes beyond the scope of this paper to investigate how parents can potentially use technology to harm or neglect children, willingly or not, I acknowledge this also deserves further attention [327]. Increasing transparency for parents over children’s online activity can sometimes be detrimental to children, and perhaps particularly so in the context of LGBTQ+ teenagers’ rights. It is therefore essential to contextualise meaningful transparency for parents in relation to the child’s development within each family and community. Moreover, not all parents are e-literate nor tech-savvy enough to conceptualise, let alone oversee, the technologies their children use, and it is not realistic to expect them all to be. Some might never play online games themselves, nor use social media such as TikTok. In 2018, 20% of UK adults were estimated to have none or limited abilities online [125]. Thus providing information over such systems can become meaningful to parents provided it is contextually appropriate information and they are interested in, able, and willing to process and act upon it.

3.7 Conclusion and main takeaways

In this chapter, I have profiled parents’ attitudes towards information that might help them understand and oversee the online technology that children use, while attempting to take the influence of their communities into account. In doing so, I contribute to address two key gaps in the transparency literature: i) the need to better define non-expert users and their specific information needs in relation to algorithmic systems; and iv) the lack of methods to evaluate (meaningful) transparency in practice (Table 1.1). Indeed, parents are responsible for protecting children online, but often lack the *meaningful* information to do so. I thus show that promoting meaningful transparency for parents over social media and online games can support digital parenting.

Towards this, I have conducted semi-structured interviews with parents of 10 to 15-year-olds who use such technologies in three communities in the UK. I have provided qualitative, empirical insights into the information needs of this relatively less-studied population, and have highlighted sources of information that might be missing (e.g. service providers), as well as key sources and types of information participants consider as helpful

(e.g. parental reviews on specific games).

To address the Literature gap 4, I have applied Nissembaun’s contextual integrity as a framework to define meaningful transparency for digital parenting and have outlined some key challenges and opportunities towards making online technology that children use meaningfully transparent for parents, e.g. by identifying norms governing the information flow between fellow parents. Overall, I show defining meaningful transparency for digital parenting is important to help parents address the many online risks children are exposed to, and to inform and guide the new legislation emerging in the field. However, we also observe in this user study with parents that non-expert users might not have a clear understanding of what transparency means. Whilst not surprising, we next look to work directly with *designers* of algorithmic systems to better describe their understanding of what constitutes transparency for non-expert users. Better accounting for the perspective of practitioners at the ‘coal-face’ is key to complement our bottom-up approach and move closer to bridging the gap between transparency theory and design practices (Chapter 2). In the next chapter, we thus explore designers’ approaches to promote transparency, alongside new standards in the field (Literature gap 2).

Chapter 4

Mind the gap: transparency guidelines & designers' practices

This chapter is based on the published paper: Bianca G. S. Schor, Chris Norval, Ellen Martin Charlesworth, Jatinder Singh. 2024. Mind The Gap: Designers and Standards on Algorithmic System Transparency for Users. In Proceedings of the ACM CHI Conference on Human Factors in Computing Systems (CHI '24), May 11–16, 2024, Honolulu, HI, USA. ACM, New York, NY, USA, 16 pages. <https://doi.org/10.1145/3613904.3642531>. I conducted all of the data collection and analysis for this project, and was assisted in the writing and editing of the paper by the co-authors Dr. Chris Norval, Ellen Charlesworth, and my supervisor Dr. Jatinder Singh.

4.1 Introduction

In this chapter, I aim to address primarily the second literature gap I have identified, i.e. the need to explore the gap between new transparency standards and design practices (Table 1.1). My aim is to help make sure designers' understanding of algorithmic system transparency is considered and better included when creating new standards in the field, as new guidelines must align with design practice to become effective. Towards this, I explore IEEE 7001's requirements and whether, according to professional designers, these can constitute helpful guidelines and metrics for algorithmic system transparency for users. To the best of my knowledge, IEEE 7001 provides one of the first set of metrics for algorithmic system transparency for users, and this study with designers is the first research published on this standard. My study thus provides critical and self-reflective insights for designers, academics, standards bodies and policy-makers willing to design and promote meaningfully transparent systems for users in practice. I first introduce this 'principles-to practice' gap.

4.1.1 A ‘Principles-to-practices’ gap

To effectively promote algorithmic system transparency, it is important to understand the perspective of designers in charge of building the systems, the challenges they are facing, and what would help them improve transparency. As discussed in Chapter 2, HCI and Responsible AI (RAI) academics call for more research on designers’ experiences to close the ‘principles-to-practices’ gap [221, 303, 358, 240]. Addressing such a gap is important to assist with building more responsible algorithmic systems going forward. As well as with other resources, standards can contribute to guiding the system development process [9]. Standards can be used by lawyers, procurement specialists, and auditors to set specifications, e.g. to increase safety and interoperability. The role of standards is acknowledged, and their recommendations widely adopted in fields such as finance and computer-aided design [353, 9]. New standards are emerging in order to set measurable levels of transparency for algorithmic systems (§4.2): published in March 2022, the Institute of Electrical and Electronics Engineers ‘Standard for Transparency of Autonomous Systems’ (IEEE 7001) [171] is among the first specifically focusing on system transparency. This standard explicitly targets designers among other audiences [171, 364] and by setting guidelines it can thus influence their design. While such standards may have the potential to promote algorithmic system transparency, I hypothesise that there is a gap between the guidance that they provide and designers’ understanding of what is helpful to increase the algorithmic transparency of their systems.

I therefore use IEEE 7001 to explore designers’ understanding of algorithmic system transparency, and the degree to which their perspectives align with the standard’s recommendations. Indeed, designers are cited among the target audience of this transparency-focused standard [364]. Like in IEEE 7001, I define designers as: “designers, developers, builders, maintainers, and operators, as well as decision-makers and procurers in organizations using and deploying autonomous systems” [171]. In line with recent HCI and RAI research [358, 79, 303, 221, 379, 248, 252], I specifically investigate designers’ perspectives in relation to new transparency standards by exploring *(i)* designers’ understanding of, and experiences building, transparent algorithmic systems; *(ii)* the degree to which their perspectives align with the standard’s recommendations; and *(iii)* some key challenges and mechanisms that can support designers alongside standards in making more transparent systems.

4.1.2 Study outline and contributions

For this, I undertake two studies with a mixed-method approach: first, I conduct an online questionnaire to survey designers’ experiences and understanding of system transparency (§4.4.1). I then interview designers about various recommendations, scenarios,

and definitions from IEEE 7001 (§4.2.4) and assess whether they find them understandable and appropriate to build transparent systems (§4.4.1). I use IEEE 7001 as an exemplar standard to explore designers’ perspectives because *(i)* it directly describes designers as part of its audience and its “measurable [and] testable” approach to transparency makes it a practical example to explore designers’ current practice (§4.2.4); *(ii)* according to its definition of designers, designers are among others directly in charge of promoting transparency in algorithmic systems [171, 364]; *(iii)* it is one of the first of few standards published on system transparency; *(iv)* it is intended as applicable across application domains (e.g. healthcare, education, transportation) and types of algorithmic systems; and *(v)* the IEEE is a prominent international professional organisation [169].

Note also that this study is not intended an assessment of the IEEE 7001 standard per se, nor do I imply such standards are to be seen as design manuals for designers. Instead, I explore current design practices and IEEE 7001 transparency recommendations for users to in order to assess and expose any gap between the two. In doing so, I contribute towards promoting *algorithmic system transparency design* by : *(i)* better understanding designers’ perspectives and experiences making transparent systems, *(ii)* describing the degree to which their perspectives align with the standard’s recommendations, *(iii)* identifying some key challenges and mechanisms that can support designers alongside standards in making more transparent algorithmic systems, and *(iv)* identifying opportunities towards developing transparency best practices.

4.2 Background

To give context, I first explore the role of designers for implementing algorithmic system transparency. I then provide an overview of recent standards in human-centred design and describe IEEE 7001 in more detail.

4.2.1 Implementing transparency for users: the role of designers

Mere calls for more transparency have been shown to be inefficient [381, 121]. Designers thus have a key role in ensuring transparency is implemented in a contextually appropriate manner. Corbett and Denton [79] call for ‘more research that moves “beyond abstractions and formalisms to drill down into the specific legal, institutional, historical, political, and cultural contexts. ...” where transparency is being applied.’ Indeed, transparency mechanisms in algorithmic systems can be ineffective or underused in practice [213, 108, 384, 280]. To close the gap between explainable AI and users’ transparency needs, Liao et al. [213] take a designer-centred approach and interview UX and design practitioners working on AI systems to understand current practices. Yildirim et al. [379] take a similar approach to understand practitioners’ perceptions around human-AI guidelines, in

particular Google’s ‘People + AI Guidebook,’ which includes a chapter on explainability [270]. Likewise, Wang et al. [358] stress that UX design practice has evolved in relation to growing calls for Responsible AI, and that such emerging practices and associated challenges deserve further academic attention. My study builds on research that focuses on designers’ perspectives [172, 292], by exploring designers’ understanding of building transparent algorithmic systems for users in relation to a new standard in the field.

4.2.2 Standards for human-centred design and transparent algorithmic systems

Human-centred AI has been increasingly studied by HCI and AI researchers [60, 286, 2, 111]. Ehsan et al. emphasise the need to operationalise human-centred perspective in XAI “to produce actionable frameworks, transferable evaluation methods, concrete design guidelines, and articulate a coordinated research agenda for XAI” [110, 112]. In this context, standards are emerging that provide overarching guidance for building more transparent systems, and thus deserve attention. The role of standards is primarily to guide stakeholders in adopting what is considered best practices, usually with general directions and requirements [353, 63]. Standards are often derived from or respond to innovative technology, for example in manufacturing and computer hardware, but can also influence the development of technology [9]. Guidelines and directives are published by a number of official standards bodies, such as the International Organization for Standardization (ISO), the International Electrotechnical Commission (IEC), and the British Standards Institute (BSI) [333, 332, 175]. Among them, IEEE has a global academic and technical community [169].

Standards are emerging in relation to human-centred design, and increasingly encompass systems’ socio-technical contexts, as well as ethical considerations [176, 168]. In November 2022, the AI Standards Hub [164] listed nine published and five pre-draft standards in relation to human-computer interaction and human-centred design in its database. ISO/IEC 30150-1 ‘Information technology. Affective computing user interface (AUI) - Model’ for example provides “a systematically defined model for affective computing user interfaces (AUI) and topics for AUI standardization” [174]. The IEEE also launched a series of ‘human standards’ focusing on the ethics of autonomous systems [63].

Some recent standards explicitly mention designers’ role in building transparent systems, in particular IEEE 7001 [364], but few have been finalised and published yet. The UK’s key body upholding information rights (the Information Commissioner’s Office) and the UK’s national institute for Data Science and AI (the Alan Turing Institute) have published guidance, aiming to provide “practical advice to help explain the processes, services, and decisions delivered or assisted by AI, to the individuals affected by them” [262]. The pre-

draft ISO/IEC AWI 12792 ‘Information technology—Artificial intelligence—Transparency taxonomy of AI systems’ likewise provides a taxonomy of information elements to help identify transparency needs in AI systems [177]. Published in November 2021 by the Central Digital and Data Office in the UK, the ‘Algorithmic Transparency Standard’ is one of the first published standards promoting transparency specifically, alongside IEEE 7001 [94]. IEEE 7001 is particularly relevant as one of the first international and umbrella standards that spans all algorithmic systems, whilst outlining specific and measurable levels of transparency for users of such systems. Moreover, IEEE 7001 explicitly mentions designers as part of its audience [171, 364]. I therefore use it to probe designers about their understanding of how to design more transparent systems. I now describe IEEE 7001 and its transparency recommendations for users before using them to explore designers’ understanding of transparency in the following sections.

4.2.3 IEEE 7001 as an exemplar standard on system transparency

IEEE 7001 is a prime example of standards on algorithmic system transparency, and one of the first to directly cite designers as responsible for developing transparent algorithmic systems. IEEE 7001 is therefore clearly relevant for the design community, and might also influence future standards, as it is indeed conceived to allow more domain-specific standards to complement it, e.g. in clinical AI [364]. In this chapter, I focus on the designers’ perspective given they are explicitly targeted by IEEE 7001, though note that they constitute but one group among many that could directly use and benefit from the standard [171, 364]. Because designers are cited as part of its audience, IEEE 7001 enables us to probe their understanding of algorithmic system transparency in relation to its recommendations. IEEE 7001 provides overarching guidelines for virtually all algorithmic systems, across application domains, and also has a potential resonance with the global professional IEEE community, i.e one of the largest technical organisation for engineering, computing, and technology information worldwide [169]. Further, IEEE 7001 enables the evaluation of transparency for (end) users, its recommendations categorised by stakeholder group. While the standard was published in March 2022, I used the August 2021 P7001 draft, as available on the IEEE website in September 2021, the studies taking place between Sept-Dec 2021. There were no significant changes between the draft and the published standard relevant for this work [170, 171].

4.2.4 IEEE 7001’s “measurable, testable levels of transparency” for users (TLs)

One of IEEE 7001’s key contributions is the introduction of six transparency ‘levels’ for users of autonomous systems, from 0 to 5, to standardise and audit system transparency at scale. Each level *“is a requirement, expressed as a qualitative property of the system which must be met,”* level 5 representing supposedly the greatest level of system transparency of all six levels. Since *“[all] levels are judged to be technically feasible while each successive level is typically more challenging,”* this study tests them in relation to designers’ experiences and practice, as they are core recommendations set out by IEEE 7001. I probe designers about the ‘**transparency levels for users’ (TLs)** as described in Table 4.1. To discuss designers’ understanding of the TLs, I use the IEEE 7001 definitions of ‘transparency,’ ‘non-expert users’ (users), and ‘autonomous systems’ (Table 4.2). I next detail my mixed-method approach.

Table 4.1: Transparency levels for users (TLs)

TLs	Definition
TL0	“barely or no transparency”
TL1	“accessible information including scenarios and general principles of operation”
TL2	“interactive training material”
TL3	“a functionality to get a brief and immediate explanation of the system’s most recent activity”
TL4	“a functionality to get a brief and immediate explanation of the system’s activity in a given situation”
TL5	“a continuous explanation of the behaviour, which adapts to the user’s needs and context”

Table 4.2: Definitions of non-expert users (users), transparency, and autonomous systems based on IEEE 7001

Key term	Definition
Non-Expert Users	“persons who have only a brief interaction or who interact every day with an autonomous system.”
Transparency	“a transfer of information from an autonomous system or its designers to a stakeholder that is truthful; contains information relevant to the causes of some action, decision, or behavior; and is presented at a level of abstraction and in a form meaningful to the stakeholder.”
Autonomous System	“a system that has the capacity to make decisions itself in response to some input data or stimulus with a varying degree of human oversight or intervention depending on the system’s level of autonomy.”

4.3 Methods: Exploring Standards & Designers’ approaches to System Transparency

As discussed, the question of how to effectively improve transparency for users of algorithmic systems remains a challenge. Here I describe my mixed-methods approach to identify support mechanisms for designers to develop more transparent algorithmic systems for users.

4.3.1 Mixed-method study design

I explore the potential gap between design practice and standards promoting transparency by investigating *(i)* designers’ understanding and experiences building transparent systems (Study 1), *(ii)* to what degree their perspectives align with the standard’s recommendations (Study 2), and *(iii)* some key challenges and mechanisms that can support designers in improving system transparency alongside standards (Study 1 and 2). The first study (Study 1) surveys designers’ approaches to design for system transparency with an online questionnaire, while the follow-up interviews (Study 2) test their understanding and opinions of IEEE 7001 transparency recommendations for users. In doing so, I do not assess how standards ultimately impact users – i.e. those using the algorithmic system. Rather, I explore designers’ perspective on designing transparency for users in relation to these guidelines. Although IEEE 7001 does not provide granular design guidance nor advice on how to apply its recommendations, I use two scenarios from its appendices as examples of when the standard can be applied for system transparency specification [170]. These enable us to better contextualise the discussions with interviewees around specific systems and transparency issues. No major ethical risk was identified for participants taking part in this mixed-method study due to its focus on professional experience and all personal data was immediately anonymised.

4.3.2 Participant recruitment: designers as one group among IEEE 7001’s target audience

I focus here on designers as one stakeholder group explicitly cited as the target audience of IEEE 7001: “The target audience of this standard are those designers, developers, builders, maintainers, and operators, as well as decision-makers and procurers in organizations using and deploying autonomous systems (*collectively, “designers”*) of autonomous systems who either wish to or are required to engineer systems that have a certain degree of transparency” (emphasis added) [171]. I recruited survey and interview participants in line with this IEEE 7001 broad definition of ‘designers,’ by advertising the studies through entrepreneurial, design, academic, non-profit, and professional networks, via LinkedIn,

previous colleagues, and direct invitations, using the following criteria, as self-described by participants: *(i)* the industry, size, and type of organisation, *(ii)* their main role in designing algorithmic systems, as they describe it, *(iii)* their gender and experience in this role, *(iv)* their main country of professional activity. Moreover, participants were involved in designing systems that fall under the standard’s definition of “autonomous systems” (see B.1), where *(i)* the standard is meant to encapsulate a wide range of autonomous systems, including systems with relatively little autonomy [171, 364] and *(ii)* participants self-referred to themselves as building autonomous systems (see examples below).

I collected 50 questionnaire responses (10 of which were discarded as they only answered demographic questions) and interviewed 22 participants among these (see Table 4.3). Out

Table 4.3: Interviewees’ technology areas, self-described roles, gender, years of experience, organisation sizes and countries of activity. Interviewees are referred as I[number] in the following sections.

ID	Technology Area	Self-Described Role	Gender	Experience (years)	Company size	Country
I1	Banking	Business Entrepreneur	Man	4.5	under 50	France
I2	Technology Company	Software Engineer	Man	30	over 100,000	USA
I3	Finance	Software Engineer	Man	15	under 50	France
I4	Recruiting & HR	Software Engineer	Man	5	under 50	France
I5	Academia	UX Designer	Woman	7	N/A	Poland
I6	Technology Company	Business Entrepreneur	Man	2	under 50	France
I7	Design	Software Engineer	Man	7	under 1,000	UK
I8	Recruiting & HR	Product Manager	Man	7	under 50	Spain
I9	Healthcare	Software Engineer	Man	10	under 50	UK
I10	Design	UX/UI Designer	Man	3	under 50	France
I11	Online Games	Software Engineer	Woman	1.5	under 1,000	France
I12	Technology Company	Computer Scientist	Man	3	under 50	Spain
I13	Education	Business Entrepreneur	Woman	2	under 50	France
I14	Finance	Product Manager	Man	3	under 100	France
I15	Finance	UX Designer	Woman	1	under 50	Israel
I16	Consultancy	UX Designer	Woman	4.5	under 100	France
I17	Consultancy	Product Manager	Man	6	under 50	France
I18	Technology Company	Product Manager	Man	9	under 50	France
I19	Public Sector	UX Designer	Man	4.5	N/A	France
I20	Academia	Data Scientist	Man	2	N/A	UK
I21	Technology Company	Hardware Engineer	Man	4	under 50	France
I22	Technology Company	Data Scientist	Man	7	under 10,000	Germany

of n=40 surveyed participants, 8 identify as developers, 8 product/business managers, 7 designers, 6 engineers/technical managers, 5 academics and 5 entrepreneurs (n=1 did not answer). These categories coincide broadly with IEEE 7001 definition of ‘designers’. Most work in technology companies (n=20), mainly in Europe (n=26) or the UK (n=8), and declare their main responsibilities are to design (n=25) or develop/build (n=17) systems (see Fig. B.1 in B.3). Out of n=22 interviewees, 13 work in tech startups and small businesses, 3 in public sector or academia, 3 in tech consultancy, and 3 in finance. 10 are engineers or data scientists, 9 are UX/UI or product designers, and 3 are business entrepreneurs. Examples of systems developed by surveyed participants include: glucose prediction based on patient (glucometer), automatic soil analysis interpretation, and

content moderation systems, and interviewees also mention having experience developing among others: mobile gaming apps, IoT devices including audio or visual recognition, and ML-based hiring and education services. Such systems fall under the IEEE 7001 definition of autonomous systems (see B.1). Note that I do not claim to represent the perspective of all possible users of IEEE 7001 with this sample, but rather to illustrate some of the themes and challenges raised by designers in charge of various aspects of algorithmic systems. Participants were offered into a randomised draw for two online vouchers as compensation, one for each study.

4.3.3 Study one (online survey): designers’ understanding and experiences of system transparency

I first use an online questionnaire to explore designers’ understanding and experiences building transparent systems for users, as well as some key challenges they might be facing and mechanisms to support them. The 48-question questionnaire (see Supplementary Material) probes participants about: *(i)* their understanding of system transparency for users, *(ii)* their experiences, approaches, and the challenges they have met to design system transparency for users, and lastly, *(iii)* their understanding of the transparency levels (TLs), how to design them in practice, and the TLs’ potential impact in system design. I randomise the TLs being considered whenever appropriate. Open text box questions also include contextual information on each TL, copied from IEEE 7001. All questions are optional and some are conditional, i.e. only appear to respondents who have selected specific answers. While all participants are asked about all TLs, each participant is only asked, at random, about how to design: *(i)* either TL1 and 3 (respectively “accessible information including scenarios and general principles of operation” and “a functionality to get a brief and immediate explanation of the system’s most recent activity”) or *(ii)* TL2, 4, and 5 (i.e. “interactive training material,” “a functionality to get a brief and immediate explanation of the system’s activity in a given situation,” and “a continuous explanation of the behavior, which adapts to the user’s needs and context”). Thus, no participant was shown the total 48 questions. This is to enable a wider exploration of issues, while limiting the length and cognitive load for participants. A survey response is taken into account in my analysis when at least one non-demographic question has been answered.

4.3.4 Study two (interviews): exploring designers’ opinions on transparency levels for users in context

To further investigate the extent to which designers’ perspectives align with IEEE 7001 recommendations, as well as some key challenges and mechanisms to increase transparency for users, I conducted follow-up individual interviews of ~30-45 mins on Zoom video

conferencing software [387] with all 22 survey participants who agreed to be contacted (see Table 4.3). The interviews started with demographic questions and a discussion of the terms ‘transparency’ and ‘standards,’ as well as the potential challenges faced by interviewees when designing system transparency for users. I then suggested an exercise with two scenarios extracted from IEEE 7001 and tested designers’ understanding and opinions about TL1, 2 and 3 in these contexts. The scenarios correspond with minimal alteration to the standard’s appendixes titled ‘Content Moderation for AI’ (*Moderation*) and ‘Credit Scoring System’ (*Credit*) (see B.2). I use the Moderation scenario to ask participants how they would improve the transparency of a video hosting website about its content moderation for content creators; the Credit scenario is about “communicating more transparently the decision-making processes to loan applicants” regarding a credit scoring system used by a loans company. I chose these scenarios for the potential risks on users they exemplify, for being quickly understandable, and for being significantly distinct. I did not test TL0, 4 and 5 in the interviews as IEEE 7001 do not recommend these TLs for users in these two scenarios. Eight interviews were conducted in English, fourteen in French – to minimise the language barrier.

Each participant was presented a single scenario and a single TL. Twelve interviewees were presented with the scenario then the TL, and the remaining ten interviewees heard about the TL before the scenario. This enabled us to test whether individual TLs were deemed more appropriate in the context of the two scenarios when presented *before* the interviewee could reflect on potential interventions in that context. Each interview consisted of three parts, with a slide deck to introduce the TLs and scenarios. The latter provide context for participants to identify interventions they think could improve system transparency, and to test their opinion regarding the relevance of specific TLs towards this. I then asked interviewees whether the scenario’s description on the slide was clear to them and could help them improve transparency for users. This allowed us to test the relevance of the TLs with no major misunderstandings on the scenarios themselves. I finally introduced one TL (TL1, 2, or 3), and asked participants whether this was clear and/or appropriate to promote more transparent systems in this scenario.

4.3.5 Data analysis

I analysed the survey open text box answers (§4.4.1) and the interviews separately with Braun and Clarke’s thematic analysis, a six phase method to identify core themes in the designers’ written comments and speech [45]. The six phases (familiarising with the data, generating initial codes, searching for themes, reviewing themes, defining and naming themes, producing the report) were completed over four steps and two iterations. I first extracted codes (phases 1 and 2) from the interview transcripts and notes. I then watched the interview video recordings, and reviewed the survey answers to group the

codes into main themes (phase 3). I finally compared and merged the themes from both studies (phase 4), before repeating the process. These are presented in Table 4.4 and the result section include numerical accounts of the key points raised by participants as well as detailed quotes illustrating these (phases 5 and 6). For clarity, I have grouped these into ‘overarching categories’ in Table 4.4. I counted as a theme any topic raised by more than three participants in either studies. I used Qualtrics [282] to deliver and analyse the questionnaire, and Otter.ai [268] for interview transcripts.

4.4 Results

To promote transparency in practice, I explore designers’ understanding of algorithmic system transparency, and the degree to which their perspectives align with IEEE 7001 recommendations. Towards this, I survey participants’ experiences and interview them to probe their opinions on how appropriate these recommendations are in two scenarios provided by the standard. I sampled participants in line with IEEE 7001 definition of designers and the mixed-method study provided consistent results overall: the gap between IEEE 7001 and designers’ perspective is important. I now describe this gap and some key challenges and mechanisms to promote transparency in practice.

4.4.1 Highlights from the findings

Study 1 explores designers’ understanding and experiences making transparent systems for users. Table 4.4 summarises the key themes raised by 40 surveyed designers and 22 interviewees, according to my thematic analysis of open text box answers (survey) and speech (interviews). I elaborate on these themes in the following subsections. The survey shows $n=30$ out of 36 agree that designing transparency for users is important ($n=4$ say they neither agree nor disagree and $n=2$ somewhat disagree with the statement). Further, $n=23$ out of 36 think it is difficult to do so, which suggests a need for supporting designers towards this. We specify in the text and each table how many participants answered the relevant survey questions (see Tables 4.6 and 4.8). Any questionnaire including at least one answer to a non-demographic question has been included.

Study 2 then investigates the extent to which designers’ perspectives align with IEEE 7001 guidelines. The 22 follow-up interviews reveal the gap between IEEE 7001 and designers’ approaches to system transparency is important: despite designers understanding the scenarios overall, they do not find the individual TLs recommended in IEEE 7001 appropriate in such contexts. Table 4.5 represents the results from each interview as a pair of one circle and one square, with a green-to-red colour scale. The large prevalence of orange and red squares highlights that only four designers found TL1-3 appropriate (which entail respectively “accessible information including scenarios and general principles of operation,” “interactive training material,” and “a functionality to get a brief and

Table 4.4: Summary of the key themes raised by designers in the survey (Study 1) and interviews (Study 2)

Overarching category	Themes
What are standards?	provide guidelines; set expectations; rarely used; GDPR
Which stakeholders?	providers; end-users
What systems & risks?	autonomous; opacity; privacy risks
What is effective system transparency?	what is “good” transparency?; which granularity?; good UX; how to define it?; what for? rarely think of it; data protection; context-specific; difficult to implement
What are the key challenges?	clash with other priorities; lack of resources; design challenges; technical challenges; defining the terms; lack of awareness; lack of incentives
What design interventions can help?	communication; legal solutions; design solutions; organisational solutions; technical solutions; empathy

immediate explanation of the system’s most recent activity”), whereas 50% found them appropriate only under certain conditions, and 32% did not. This suggests that interviewees are unlikely to implement these IEEE 7001 recommendations in similar contexts; and again, that more support seems needed to bridge the standard and designers’ approaches in practice. Moreover, the interviews confirm participants face multiple challenges in implementing algorithmic system transparency for users, and yet they are able to suggest design principles towards it.

Scenarios & Transparency Levels (TLs)	TL1		TL2		TL3		Total nb of interviews
Scenario ‘Moderation’ (scenario then TL)							6
Scenario ‘Moderation’ (TL then scenario)							5
Scenario ‘Credit’ (scenario then TL)							6
Scenario ‘Credit’ (TL then scenario)							5

In each cell, 1 interview with 1 designer is depicted by a pair of 1 circle and 1 square

Scenario clear to interviewed designer
 Scenario rather clear
 Scenario unclear
 TL relevant according to interviewed designer
 TL relevant only under certain conditions
 TL not relevant

Table 4.5: A Summary of the interview results for the Moderation and Credit scenarios and TL1-3. Two combinations (TL2 then Credit scenario; TL3 then Moderation scenario) were tested once due to time constraints.

4.4.2 Few participants know of or use standards to design user-facing algorithmic systems

In addition to investigating designers’ understanding of IEEE 7001, I probe them on their general understanding and use of standards. Almost none of the 22 interviewees knows of or uses standards for the user-facing aspects of algorithmic system design. Eighteen do not use standards to design user-facing algorithmic systems and ten do not seem to know what

an official standard is (“I don’t know what standards are” (I12), “What do you mean?” (translated, I6), “I don’t know any standards” (translated, I14)) or cite the EU General Data Protection Regulation (GDPR) as an example of a ‘standard’ which is actually a regulation (though it does encourage the use of standards and best practices). None cite standards specifically on algorithmic system transparency, though note that these were still under development and thus not widely accessible when I interviewed them in late 2021. While this is not surprising—given not all standards target designers directly—this indicates the challenge for IEEE 7001 to impact real-world design practices. Indeed, many participants cite other resources they use to design user interfaces, such as internal guidelines and processes, or external ones such as the Nielsen Norman group’s UX best practices (3 interviewees) [145]. These seem more widespread among designers, and could represent a pathway towards improving system transparency. Standards bodies could thus better engage with designers to raise awareness of standards and shape their specifics.

4.4.3 IEEE 7001 transparency recommendations are critiqued by participants

While participants do not currently use standards to design user-facing algorithmic systems, they also question the relevance of IEEE 7001 transparency recommendations for users.

4.4.3.1 Few participants say they are likely to implement the TLs.

Only 10 out of 40 surveyed designers declare being likely to provide TL2 (“interactive training material”), TL3 (“a functionality to get a brief and immediate explanation of the system’s most recent activity”), and TL5 (“a continuous explanation of the behaviour, which adapts to the user’s needs and context”), and n=14 to provide TL4 (“a functionality to get a brief and immediate explanation of the system’s activity in a given situation”) going forward. On probing whether the TLs are currently implemented, n=17 say TL1 (“accessible information including scenarios and general principles of operation”) is provided half of the time, whilst the majority (between n=21 and n=29 participants think that TL1 to TL5 are rarely implemented). This shows IEEE 7001’s impact on systems design could be limited in practice. The following subsections highlight why participants question how appropriate IEEE 7001 recommendations are in a given context.

4.4.3.2 IEEE 7001’s scenarios and TLs are overall clear but not precise enough to enable actual implementation.

Both the Moderation and Credit scenarios provided by IEEE 7001 seem overall clear for interviewees, though not for all (Table 4.5). As discussed in §4.3, these provide examples of when IEEE 7001 recommends TL1-3 for users as system specification. I use the scenarios

as contexts to gather designers’ opinions on how appropriate TLs’ are for users, and thus need to confirm interviewees understand the scenarios and TLs. Two participants find the Moderation scenario rather unclear: “It feels very heavy, vague” (I15), “how do I intervene in this?” (translated, I10). Another explicitly asks us to explain it to them. Likewise, one participant describes the Credit scenario by saying: “based on my experience writing requirements, I would say it’s badly written, go back and try again” (I2), which suggests the wording could be clarified. Similarly, TL1-3 are considered clear, though three interviewees find TL2 unclear (“I don’t understand what that means. Can you give me an example?” (I15)), and two say TL3 is too abstract: “this is basically saying the solution to transparency in the system is to explain to the user in simple language why a decision has been made in the system, which makes sense, but that’s incredibly abstract, (...) what does that system look like? What is the interface? What the actions are that the user does and at what point? (...) what this looks like visually is still not clear.” (I7).

Indeed, while participants seem to understand IEEE 7001 scenarios and TLs overall, over half of them think the three TLs do not provide enough information to enable designers to implement them in practice. When asked about TL1, one interviewee points out how unclear it can be to evaluate whether it has been effectively implemented: “there is no information here about what kind of users they are, what things they might need to know, what things they might want to know, how they would best receive information. Some people like to see videos, some people like to read texts. Some people like to be told. Some people like to be shown. Is the user disabled in anyway? There is not enough specificity here to provide ... Have I provided accessible information [i.e TL1], yes or no? Nope, can’t answer that question” (I2). Similarly, another participant wonders about TL3: “but which granularity do you go for?” (translated, I21). I acknowledge that many standards are not meant to provide granular design guidance, but their target audience (here designers among others) should still be able to understand what mean in practice the “measurable, testable levels of transparency” IEEE 7001 recommends they reach in their systems, and in this context my findings suggest otherwise. I do not test TL0, 4 and 5 with interviewees because they are not recommended by IEEE 7001 for these scenarios, but my survey results suggest similar findings to the interviewees. In all, IEEE 7001’s lack of precision for interviewees does not undermine their ability to discuss the TLs in the two scenarios, but is part of what they find problematic, and could hinder the standard’s adoption. Those involved in building standards should take note.

4.4.3.3 Participants seem to find TLs difficult to implement and not always appropriate in the IEEE 7001 scenarios.

Most TLs seem difficult to implement, according to surveyed designers. As a reminder, IEEE 7001 considers TL1 the easiest to implement and TL5 the more challenging. The

survey reflects this only for TL1 and 5 (see Table 4.6), e.g. it shows TL2 is deemed the most difficult to implement after TL5. Only TL1 seems ‘easy’ for the majority (n=17). Whilst

Perceived Difficulty of Transparency Levels	Total	Extremely difficult	Somewhat difficult	Neither easy nor difficult	Somewhat easy	Extremely easy	N/A
TL1: accessible information including scenarios and general principles of operation	32	0 0%	6 19%	7 22%	11 34%	6 19%	2 6%
TL2: interactive training material	32	2 6%	14 44%	6 19%	5 16%	2 6%	3 9%
TL3: a functionality to get a brief and immediate explanation of the system's most recent activity	31	4 13%	7 23%	7 23%	10 32%	1 3%	2 6%
TL4: a functionality to get a brief and immediate explanation of the system's activity in a given situation	31	4 13%	8 26%	11 35%	8 26%	0 0%	0 0%
TL5: a continuous explanation of the behaviour, which adapts to the user's needs and context	32	12 37%	14 44%	5 16%	0 0%	0 0%	1 3%

Table 4.6: Numbers of survey participants show most transparency levels (TLs) seem difficult to provide.

IEEE 7001 provides the Moderation and Credit scenarios as examples where TL1-3 should be implemented, I observe that out of the 11 designers interviewed with the Moderation scenario, the majority is not convinced that TL1-3 are appropriate here (Table 4.5): three say the TLs could be potentially appropriate, but only if designed in a specific way. For instance, one says about TL2: “the important thing is to not make it too heavy because users don’t actually care about that, like they just want to get through it, they’re not in the mood to read or do a heavy on-boarding, so yeah keep it short and actually useful” (I15). Likewise, two participants say the TLs are not the best options, and another explains TL3 is not appropriate in the Moderation scenario, as it can be “frustrating” for users: “I think having the information afterwards would often just be frustrating. I think it would be like a step up from where we got to but, often like, to get a decision, you have to put in a lot of time, and it’s time you could save like if you had more information about what it was actually looking for in the first place. Maybe like insurance, it would cause more people to game the system, which isn’t good, like to lie, but the content creation is like, you could literally decide to create a different video.” (I20).

The same TLs are judged even less appropriate in the Credit scenario: all interviewees find they are either incomplete, unrealistic, or not adapted to this socio-technical context. One comments about TL2: “I don’t really see the point of designing a simulator [for users]” (translated, I16). Another describes TL1 as likely “frustrating” for users: “there is clearly a need to explain to the users why a loan has been denied (...) to me, what’s missing here is what to do, the next steps, how to improve my credit scoring, can I do something to change my situation? Because only presenting information can create frustration (...) it’s essential to be able to act upon that information, otherwise it’s just a justification from the provider, from the company who is backing itself” (translated, I14). Another comments “I think [TL1] is clear, I would question its usefulness,” before highlighting the risk it could lead to for users: “assume that you are writing this for a malicious software engineer, who will take this specification and implement it to the letter in the most unpleasant way (...)

Suppose I had three or four reasons for explaining why I’m going to say no to a loan, and I choose to pick the one that is going to cause the most offence” (I2). Finally, one participant finds TL3 “pointless” but TL2 more appropriate, as it provides information to the user before they start using the system: “if the goal is transparency [on the use of personal data], I don’t understand the point of telling the user after they have given access to their personal data. For me, it should happen beforehand” (translated, I18). These results were consistent regardless of the order in which the scenario and the TL were presented. Thus they outline the challenges of creating and implementing transparency standards, given TLs are central to IEEE 7001 guidelines and TL1-3 are recommended in these scenarios.

4.4.4 Key challenges for improving system transparency

4.4.4.1 What ‘transparency’ actually means is unclear to most participants.

Most participants expressed that they find the term ‘transparency’ unclear. Two designers differentiate transparency between hidden and visible features, whilst three consider ‘transparency’ good UX and clear communication. For example, one participant stresses the challenge of explaining complex machine learning algorithms to users intelligibly and faithfully: “the first thing I think about is the algorithm, the second thing is the educational message. Transparency and simplification enable to communicate a complex message to non-expert users (...) but you have to take shortcuts for it to be impactful and digestible and so, as a result, you lose on transparency, paradoxically” (translated, I14). Another acknowledges there are different types of transparency, and cites the UX of an order process vs. the system’s ethical/environmental impact: “it depends what transparency we’re talking about, like is Amazon transparent on the ethical and environmental implications of your interaction with the system? No, definitely not, no e-commerce website or very few would be. Is it transparent on the behaviour of the system? For example, once you’ve interacted with it to place an order, what the status of that order is? What the journey of that order is? How you can interact with that order? Yes, in that sense it is very transparent” (I7). After being given the IEEE 7001 definition of ‘transparency’ (§4.2) in the survey, one interviewee expressed that they found this definition “incomprehensible” (I2), and several others asked what system transparency means, which confirms the general confusion regarding the meaning and significance of term in this context. Beyond the quality of IEEE 7001 definition, this suggests a wider community understanding of what transparency means in practice is needed.

4.4.4.2 Participants’ awareness about the risks posed by algorithmic systems seems primarily focused on privacy.

Most surveyed designers think they are aware of and identify risks posed by their algorithmic systems: n=33 identify their systems as directly impacting users in high-stakes situations, e.g. in employment, healthcare, credit risk etc. Among the risks posed by their systems, the privacy risk for users and their personal data is identified the most often (n=19), followed by interpretability risks (n=13), i.e. complex model that may not be fully understood by users. There is a general awareness of the risk of gathering personal data, perhaps due to GDPR’s presence in Europe: when asked what information they provide to users to mitigate such risks, almost half of the respondents say they communicate GDPR-compliant information or privacy terms (see Table 4.4). My study thus indicates a similar push for change on other issues linked to algorithmic system transparency is still needed.

4.4.4.3 Participants identify several challenges to design transparency: clash of priorities, lack of resources, and clear definitions.

Moreover, designers identified obstacles for system transparency (Table 4.4, Table 4.7). The key challenges from surveyed respondents were potential clashes with other priorities

Table 4.7: Non-exhaustive list of additional challenges in designing for transparency, as raised by interviewees.

Other challenges	Examples cited
user or product-specific	fear of using technology
difficult choice of wording and format	complex concepts are difficult to explain
users’ limited attention span	many users skip onboarding
lack of specification to improve transparency	misunderstandings on what transparent interfaces mean
lack of empathy/awareness among internal stakeholders	users’ needs are not always prioritised
quantifying the added value of designing for transparency	lack of clear metrics to convince hierarchy

(including commercial ones) (n=6), followed by the lack of financial means (n=3). Some mention the technical challenge of providing explanation for automated decisions, and the design challenges of “understanding the end-users difficulties” and “getting the right visual design.” When asked about TLs specifically (Table 4.8), designers raise the lack of time, a clash with other priorities, the lack of designing tools and methods, and the difficulty of defining terms, e.g. ‘transparency’. The interviews confirm this: almost all designers raise at least two challenges to implement transparency in user-facing algorithmic systems. Many match those identified in the survey (Table 4.8), others are summarised in Table 4.7.

Table 4.8: Most surveyed designers raise challenges to implement IEEE 7001 Transparency Levels (TLs).

Challenges to implementing TLs	Nb of respondents
Lack of time	24 (60%)
Clash with other priorities	18 (45%)
Lack of designing tools and methods	13 (32%)
Defining the terms (e.g. transparency)	12 (30%)
Lack of awareness in teams	9 (22%)
Lack of financial means	9 (22%)
Lack of awareness in management	8 (20%)
Other	7 (17%)

4.4.5 Support mechanisms for designers: raising awareness, providing tools, and outlining best practices

These challenges represent opportunities as to ways forward on supporting meaningful transparency for users; indeed, participants then went on to identify support mechanisms and design principles to promote transparency in practice, as detailed below.

4.4.5.1 Raising designers’ awareness is needed.

My study reveals the need and potential for raising awareness about transparency among designers. Though seven designers say they had never thought about transparency before, many show further interest in transparency after participating in the study: “it makes me want to go and have a look at a couple of websites to see if there is this notion of transparency there and how it materialises in practice, I don’t feel that it is, to put it simply, so I am going to have a look, actually, straight after this” (translated, I21). Half of the participants say they feel curious or more aware about transparency issues in the interviews (“it awakens me to this topic that I do not see from this angle” (translated, I18), “it’s made me think about what we mean by transparency” (I7)), and two even shared transparency-related news and resources via email after the interview (including a screenshot of YouTube’s change of general conditions, where a paragraph about ‘transparency’ had been circled by the designer). Three participants also raise doubts about whether making systems more transparent for users is part of their role, or more broadly raise the question of who is responsible for it, whereas others state that all the stakeholders should contribute to implementing system transparency for users, (“that would be the job of the UX designer (...) maybe in tandem with the product manager?” (I7)) and that transparency considerations are relevant for all aspects of system design: “it’s always interesting because then you can take these principles, you can apply them in different aspects, whether this is designing of a web page, designing of a, I don’t know, larger system, designing of some

back end code” (I9). Thus, further research toward raising designers’ awareness of the importance of designing for transparency, like during the interview, as well as incentivising them, and better defining the key stakeholders responsible for developing, maintaining, auditing or regulating it is crucial for improving system transparency. Similarly, raising awareness about new transparency-focused standards could make such standards more effective in practice given these are still recent and many participants declare not knowing or using standards to design user-facing algorithmic systems (§4.4.2).

4.4.5.2 Several designers identify help that could support them in designing for transparency.

To overcome the challenges raised above, I asked participants to describe what resources could be useful for them to improve system transparency for users. When asked to advise others trying to design for transparency, only seven surveyed designers recommend using standards; they primarily recommend asking more experienced colleagues (n=14) or finding other sources of help (n=10) (see examples below). Some also comment on the need to “work with users to determine what they want to know and what is a suitable format to present that information.” I discuss calls for co-creation below. Others emphasise the need to “understand requirements and constraints in the business context,” “use empathy” or “community resources,” better communicate or “hire an analyst to interpret the model.” Moreover, surveyed designers explicitly comment on what could help them to improve system interfaces in practice: they need “examples and best practices,” as well as “explanations on how and when to implement them,” but also “targets on what is necessary and when; what the end user is in need of; what the company I work for actually wants to provide.” Similarly, “a clear description of why we do it, a clear demand from end user” would help. A few respondents also wish for more resources: either more “understanding from [their] management” and of the user’s needs, “more time,” or “a person dedicated to solving issues [related to algorithmic system transparency].” Lastly, better tools were often cited as required, e.g. frameworks for “quick and easy implementation” to improve system transparency. These requirements confirm the challenges raised above, but also pave the way for standards and research to better support designers in designing for transparency.

4.4.5.3 Participants have suggestions for tools that they think can assist in making systems more transparent.

Participants are able to identify tools that they think could help increase system transparency for users. Table 4.9 summarises the tools mentioned in the survey. The interviews confirm these, as all participants suggested interventions they thought might facilitate transparency in algorithmic systems. Given participants’ initial uncertainties regarding what transparency means in this context, not all may be relevant, which confirms the

Table 4.9: Tooling suggested by surveyed designers to improve system transparency

Category of tools	Examples cited
user-centric tools	user tests, shadowing, workshops, FAQ, communication material, tutorials, videos, user guides, explanations
internal tools	guidelines, training for employees, writing tools to simplify language, examples of successful approaches
technical features	functional specifications, technical release notes, and user guides
legal tools	audit records, online dispute resolution, contracts, memos on GDPR, GDPR guidelines

need to raise designers’ awareness about transparency. Yet, the findings indicate that, when given an appropriate understanding of transparency, designers are likely to suggest interventions they think may increase transparency. This is encouraging as it shows the potential for productive collaborations with designers such as my participants and those producing standards and guidance. Thus supporting them in identifying best practices can offer an opportunity to improve algorithmic systems (§4.5.1).

4.4.6 Principles for ‘good’ transparency in algorithmic systems for users

I further asked surveyed designers as to what transparent user-facing systems might consist of. Table 4.10 lists the main principles designers identified in the survey’s open text boxes. Whilst non-exhaustive nor unanimous, this list summarises themes raised at least twice. For designers, ‘good’ systems provide information this is, for example, simple, intuitive, playful, and given at the right moment in the user journey – see Chapter 5 for a further investigation of this later point. Though further assessment of these principles is needed, it can be helpful for identifying best practices and examples of more transparent systems. Note that these recommendations (Table 4.10) do not match those provided by IEEE 7001 as the standard is not intended to provide granular design guidance. As a reminder, IEEE 7001’s main contribution is the transparency levels: TL1 “accessible information including scenarios and general principles of operation;” TL2 “interactive training material;” TL3 “a functionality to get a brief and immediate explanation of the system’s most recent activity;” TL4 “a functionality to get a brief and immediate explanation of the system’s activity in a given situation;” and TL5 “a continuous explanation of the behaviour, which adapts to the user’s needs and context”). This further indicates a gap between what designers and standards consider appropriate to promote transparency. Such ‘bottom-up’ information, as provided (and desired) by practitioners at the ‘coal-face’, is useful for those—be they practitioners, standards bodies, academics, and policy-makers—that strive towards increasing algorithmic system transparency in practice.

Table 4.10: Principles for ‘good’ transparency over algorithmic interfaces according to surveyed designers

For designers, ‘good’ systems provide information that...
* is clear and easy to understand
* is complete and not hidden on the user interface
* is accountable or certified
* gives confidence and trust to users
* is accessible and gives the correct level of information
* is simple, intuitive, playful, and empathetic
* enables to configure the system, override it, and identify who is responsible for its impact
* is compliant with personal data protection
* enables informed decision-making and interaction
* is given at the right moment in the user journey

4.5 Discussion

Mere calls for increased system transparency can be ineffective in practice [121, 381]. However, transparency does have an important role to play – it can be helpful in supporting accountability, fairness, scrutiny, contestability and other aims [360, 68, 5, 380], which work towards managing the risks of algorithmic systems. By exploring designers’ understandings and experiences regarding algorithmic system transparency as well as their opinions and perspectives on the specific transparency recommendations from IEEE 7001, I have highlighted some of the key challenges and principles for improving system transparency alongside standard guidelines. In doing so, I help draw further attention to the ‘principles-to-practices’ gap [240], showing that it extends to transparency standards, and indicate the need and opportunities for addressing this gap to help ensure that ongoing efforts (e.g. standards, RAI initiatives) are effective in practice. As one of the first international standards focusing on transparency, IEEE 7001 sets measurable levels of transparency in algorithmic systems according to specific stakeholder groups. Despite its potential for promoting transparency for algorithmic system users, my results show that participants do not find its recommendations relevant in the scenarios it provides as examples. I argue such standards should be part of a wider concerted effort to raise designers’ awareness alongside more engagement with stakeholders (i.e. standards organisations, users, designers etc). I now suggest further mechanisms to better support designers before discussing the study limitations.

4.5.1 Suggestions for supporting designers in making algorithmic systems more transparent for users

Based on the results, I now outline ways in which standards, raising transparency awareness, and closer collaborations with various stakeholders can better support designers in

improving transparency for users.

4.5.1.1 Raising awareness about transparency among designers.

My study reveals the potential impact of transparency-focused standards, such as IEEE 7001, might be limited until there is more awareness amongst designers generally (i.e. beyond those already engaged in interested communities) of the need and importance of transparency, and of the role of such standards. For example, my qualitative insights highlight how the short online interviews (study 2) raised not only awareness but also interest and curiosity for system transparency among participants (see §4.4.5.1). I encourage more efforts in this direction, for example through co-design approaches. Co-design has already been tested to help close the ‘principles-to-practices’ gap, e.g. with an impact assessment framework for responsible AI values [303] and a checklist to operationalise fairness in AI [221]. Likewise, I argue co-design [321, 241] and stakeholder engagement (‘participatory AI’ approaches [37, 390]) are means not only for improving transparency implementation (where users and other stakeholders are part of the problem-solving endeavour), but also for reaching agreements on what transparency means in practice (with closer connections between standards bodies and designers for example [379]). Indeed, even within the transparency research community, the meaning, value and implementation contexts for ‘transparency’ are still being debated [79]. In this context, the AI standards hub launched by the Alan Turing Institute, UK, organised two inter-disciplinary workshops in January 2023 for individuals to take part in developing two transparency-focused standards [7].¹ I argue such discussions are needed, including to highlight the limitations of transparency, and how it might clash with other core values of responsible AI, such as security, privacy and data protection [122, 185]. Likewise, standards should be part of a wider concerted effort to raise designers’ awareness alongside more engagement with stakeholders, as promoted by the AI standards hub [163].

4.5.1.2 Further identifying examples for future standards and transparency metrics.

My study, along with a lack of literature in the space, suggests no clear consensus has emerged on what best practices might be so far. And yet one key suggestion from the interviewees is the need to provide them with clear examples of how to design for transparency. One summarises this as follows: “a couple of examples would have been great (...) because, in fact, since I am not used to thinking about it in theoretical terms, as a result the way [TLs] were phrased, I struggled to link it back to what I might have seen,

¹Pre-drafts ‘ISO/IEC AWI TS 6254 Information technology — Artificial intelligence — Objectives and approaches for explainability of ML models and AI systems,’ and ‘ISO/IEC AWI 12792 Information technology — Artificial intelligence — Transparency taxonomy of AI systems’ [7, 177]

because I’m immediately going to ask myself: what does it look like? I understood all the words, you see, it’s just that I didn’t manage to link it back to things that I do in practice in UX or in design” (translated, I19). Thus providing more real-life examples in standards might constitute an opportunity to further align them with designers’ understanding of transparency in practice, and thus facilitate their implementation. As a result, showcasing examples of good transparency practices could become a lever to improve system designs overall. My findings regarding transparency characteristics of ‘good’ systems (Table 4.10) could potentially inform, support, and perhaps become transparency standards. Such characteristics could be used to adapt transparency recommendations and bring them closer to designers’ perspective, thus helping to operationalise more responsible AI going forward [110, 112].

4.5.1.3 Supporting designers in experimenting with transparency interventions.

As our results show, all interviewees were able to come up with design interventions they thought could increase transparency in a given system. I acknowledge their relevance might vary depending on designers’ understanding of transparency and experience. However, the design interventions recommended by IEEE 7001 do not seem to be the most effective, according to participants, in making systems more transparent. This aligns with the idea that transparency becomes *meaningful* only when contextually appropriate [79, 68]. Building upon the ideas raised in §4.4.5.3 and Table 4.10, I thus encourage more work on facilitating designers’ experimenting with transparency interventions in context. For example, designers seem to report various tools to capture, manage, and collaborate on ideas [172, 292]. These could be used to promote a collective effort to improve transparency in practice. Such design efforts can coincide with Felzmann et al.’s transparency-by-design framework “that can act as a reflection tool for different stakeholders to integrate transparency considerations into their practice” [121].

4.5.1.4 Engaging various stakeholders more in standards development.

One of the key findings from this study is the need for further engagement between various stakeholders, namely standards organisations, designers, and users on ways for making systems more transparent. I identify a gap between standards’ recommendations and designers’ approaches to improving algorithmic system transparency (see quote in §4.5.1.2). This confirms the ‘principles-to-practices’ gap identified in the responsible AI literature [221, 79, 303, 240] extends to standards bodies promoting transparency. Those developing standards could therefore better support designers by engaging with them more closely, and vice-versa, such that the standards are more targeted, relevant, and fit for purpose. This collective approach to promote transparency has also been highlighted by

Rakova et al., who show organisational structures have an impact on the implementation of responsible AI initiatives [285]. One starting point for such efforts could be to focus on addressing the main challenges faced by designers willing to build more transparent systems, such as the ones I identify in §4.4.4.3. My study also indicates more engagement with users is key to move forward [213] (see, e.g., §4.4.5.2). Moreover, several designers have raised doubts about whether making systems more transparent for users is part of their role (see §4.4.5.1). Connecting practical interventions directly with someone responsible for designing, monitoring, and/or auditing transparency over algorithmic systems and its potential impact could therefore greatly facilitate a general (and practical) understanding of transparency and improve system designs. The recent launch of the AI Standards Hub (Oct 2022) is an example of initiatives and inter-disciplinary communities that aim to bridge the gap between standards bodies, academics and practitioners [163].

4.5.2 Study limitations

Finally, I identify three areas that can further complement and confirm my findings. First, the sample size is limited due to time constraints, and influenced by the main author’s professional and online networks. For example, most participants are based in France/Europe: while 75% of surveyed designers come from France/Europe, 20% from the UK, 5% from the Middle East, Asia, North America, and Oceania, and none Africa and South America. Therefore similar studies could be conducted with a larger or different sample, for example with those from the Global South. I also acknowledge that designers are one among other stakeholders explicitly targeted by IEEE 7001. I focus on their approach and do not claim to represent the perspective of all possible users of IEEE 7001, though these deserve attention going forward. I also faced a language barrier in this study: despite speaking at least general English, most participants are not native English speakers (only 18% of interviewees are) and several raised the language barrier as a factor hindering their understanding of the terms used in the survey. I thus translated orally into French the scenarios and TLs for 5 interviewees, who reported they understood them overall. Given the global nature of transparency concerns, it is important that prominent standards are made available and accessible across a broad range of communities. Lastly, this study is not an assessment of all standards, nor of IEEE 7001 in its entirety. My aim was to better understand designers’ perspective on algorithmic system transparency by probing their understanding of IEEE 7001 recommendations for users. I chose IEEE 7001 for this study, given its high profile, overarching scope, and level of development, as a representative exemplar. I use it to indicate potential real-world issues stemming from algorithmic transparency standards, though I acknowledge other standards have been published since. IEEE 7001 is significant as it is *(i)* one of the first published international standards that seeks to address system transparency specifically, *(ii)* from IEEE which represents a large

international professional community, and *(iii)* spans virtually all algorithmic systems. This study thus paves the way for a broader assessment of IEEE 7001, and studies on other standards to compare and expand my results.

4.6 Conclusion and main takeaways

This chapter has explored the gap between standards’ recommendations and design practices (Guideline Challenge (RQ2)). It takes a step towards bridging this by better understanding designers’ approaches to build transparent systems, suggesting mechanisms to support them in improving transparency for users, and discussing considerations for developing future transparency standards (Literature gap 2). Closing this gap is important for ensuring more responsible algorithmic systems going forward. Towards this, I have explored designers’ experiences of transparency and understanding of IEEE 7001 guidelines (‘Standard for Transparency of Autonomous Systems’), i.e. one of the first international standards published in the field. Despite this standard’s potential, my mixed-method study shows that IEEE 7001 might not be effectively implemented by designers in practice, due to a mismatch between their respective approaches; e.g. interviewees do not seem to find key IEEE 7001 transparency recommended requirements relevant in the scenarios provided by the standard. Moreover, most participants seemed eager to learn more about the concept of transparency despite not always being familiar with it, so this study highlights the potential for raising awareness among designers. Whilst I acknowledge transparency is not a panacea, it plays an important role in facilitating understanding, critique, contestation and accountability, providing a means that can help in better aligning systems with the public interest.

I therefore call for more collaborative efforts across communities that *(i)* raise designers’ awareness of the importance of designing for transparency in algorithmic systems in combination with *(ii)* further engagement between stakeholders (i.e. standards organisations, designers, users etc.). I further contribute by highlighting that more research is needed to *(iii)* solidify definitions, enable a better understanding of transparency, and facilitate its implementation in algorithmic systems, and *(iv)* identify examples, best practices, and users’ needs. This chapter thus identifies important challenges, but also mechanisms to support designers in making algorithmic systems more transparent for users.

In the next chapter, I aim to explore all four of these calls by directly involving non-expert users as co-designers. I specifically study meaningful transparency in relation to HCXAI design recommendations by grounding our study in a under-studied clinical context: gynaecology. In doing so, I thus aim to further explore the Guideline Challenge (RQ2), and contribute towards operationalising such recommendations in a real-world setting (Literature gap 3). Moreover, in light of my findings, I also observe that IEEE 7001’s

measurable requirements (i.e. TLs) have not won the support of most of our participants when it comes to improving transparency for users. Consequently, these transparency levels are unlikely to be used as transparency metrics by the majority of developers today, leaving the Assessment Challenge (i.e. lack of methods to assess transparency) still unanswered. I will come back to this challenge and the related literature gap 4 in Chapter 6.

Chapter 5

Operationalising HCXAI recommendations: a co-design approach with clinicians

This chapter is based on the following paper: Bianca G. S. Schor, Emma Kallina, Jatinder Singh, Alan Blackwell. 2024. Meaningful Transparency for Clinicians: Operationalising HCXAI Research with Gynaecologists. In Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24), June 3–6, 2024, Rio de Janeiro, Brazil. ACM, New York, NY, USA, 14 pages. <https://doi.org/10.1145/3630106.3658971>. I conducted all of the data collection and analysis for this project. I was assisted in hosting and audio recording the workshop and interview in Amsterdam by Emma Kallina. I was assisted in the writing and editing of the paper by my PhD supervisors Dr. Jatinder Singh and Prof. Alan Blackwell.

5.1 Introduction

After exploring standards providing requirements for transparent algorithmic systems, I now address the Guideline Challenge (i.e. RQ2: recent guidelines and design recommendations must align with designers' perspective and users' needs) by exploring the design recommendations currently offered by the HCXAI literature. To my knowledge, this is the first study using a co-design approach to explore what constitutes meaningful transparency for clinicians with gynaecologists. In doing so, I thus contribute to answering the first and third literature gaps I have outlined: i) the need to better define non-expert users and their specific information needs in relation to algorithmic systems, and iii) the need to better operationalise design recommendations from recent HCXAI publications in real-world settings (Table 1.1). My aim is to make sure such guidance for building transparent systems is both contextually appropriate for specific users, here clinicians, and applicable

in real-life settings, here ovarian cancer diagnosis in the UK and the Netherlands. Indeed, I argue that gynaecologists constitute an under-studied group of primary non-expert users of CAD tools and thus deserve more attention. I first justify my focus on clinical AI implementation.

5.1.1 Transparency for clinical AI implementation

Safe and ethical implementations of clinical AI can bring significant benefits to our healthcare systems, e.g. by improving patient outcomes or accelerating drug discovery [350, 28, 311]. Indeed, AI research in healthcare has boomed in the last decade [14], yet real-life implementations are still facing obstacles [181, 188, 266]. *Algorithmic system transparency* can assist AI applications in such a high-risk contexts, e.g. by facilitating accountability and calibrating clinician’s trust – i.e. helping them decide how much to trust the system [384, 360, 68, 53]. I define it here as the transfer of information from a developer or a system to clinicians about the system’s design, training data, behaviour, and potential impact [171, 364].

In this chapter, I argue transparency can help bridge the gap between academic research and clinical AI implementation when *meaningful* to clinicians. Towards this, I explore recent HCXAI recommendations with clinicians in order to better operationalise them and further define meaningful transparency for clinicians. Whilst a growing number of papers study clinicians’ perception of AI systems [315, 350, 376], what meaningful transparency means to them remains unanswered [336, 31, 385].

To answer this question, I ground this study into the specific clinical context of gynaecological imagery. Using systematic reviews of the HCXAI literature, I extract design recommendations for transparent and explainable AI systems, and compare these to clinicians’ own concerns. By better understanding clinicians’ perspective on what constitutes *meaningful* transparent AI systems for them, I aim to complement and help operationalise current HCXAI recommendations.

5.1.2 Gynaecologists as users of clinical AI systems

I focus on gynaecologists as users to test and ground these HCXAI design recommendations into a specific and relatively less-studied field of practice. As highlighted by Ehsan et al. [109, 111], “understanding who interacts with the black-box of AI is just as important as “opening” it, if not more.” Clinicians in a variety of specialities have already been studied as users both of clinical AI systems, and of particular transparency interventions such as image saliency maps [51, 385]. In contrast, while gynaecologists frequently use medical imagery such as ultrasound scans (e.g. for diagnosis) [183], to date AI has had limited impact in the field [100]. Advances are likely in the near future, given current

active research in AI for gynaecological imagery [48, 162, 229, 223, 148], and efforts are being made to integrate them into clinical practice [118]. This is therefore an appropriate point for the XAI, FAccT, and HCI communities to consider the transparency needs of gynaecologists who would interact with AI systems in their clinical practice (§5.2.2).

5.1.3 Study outline and main contributions

To do so, I involve gynaecologists as co-designers to better describe meaningfully transparent AI systems for use by clinicians. As a real-life case study for co-design, I use a recent medRxiv preprint describing a deep learning model to assist in ovarian cancer diagnosis [48]. I carried out a co-design workshop (§5.2.3) with n=10 gynaecologists and interviews with n=6 gynaecologists in the UK and the Netherlands (§5.3). I show that HCXAI must better account for clinical teams with different types of gynaecologist users, and that the timeliness and relevance of the information provided about the AI-based tool is key for transparency to become meaningful for clinicians throughout the design lifecycle. Gynaecologist participants particularly ask for transparent information about the tool’s conceptualisation, role, and purpose prior to its implementation in clinical practice. My main contributions include: i) testing recommendations from the recent HCXAI literature, with a prospective, real-life AI application in a relatively less-studied clinical domain; ii) describing and visualising how gynaecologists understand meaningful transparency for clinicians; iii) outlining research and design recommendations towards meaningful transparency for clinicians; and iv) expanding HCI and AI research in women’s health by directly engaging with gynaecologists as users and co-designers. Exploring such issues is key to implement more transparent AI systems that meet clinicians’ information needs.

5.2 Background

I first describe some key publications on algorithmic system transparency in relation to clinical AI (§5.2.1), before outlining the use of AI in gynaecological imagery (§5.2.2) and co-design studies with clinicians in the HCXAI literature (§5.2.3). The HCXAI design recommendations are described in detail in the next section (§5.3.1).

5.2.1 Defining algorithmic system transparency for clinicians

Algorithmic system transparency is crucial in high-risk domains like healthcare. It can support AI implementation, e.g. by helping clinicians effectively trust AI tools [53], but also help address risks for patients by supporting accountability, i.e. allowing scrutiny and challenge to individual decisions of a system [360, 249]. Indeed, Ehsan et al. [108] suggest social transparency –i.e. an explanation of AI-mediated decision-making that

incorporates the socio-technical context— can help calibrate users’ trust in algorithmic systems and improve decision-making. Nonetheless, defining transparency for clinicians remains an open challenge. For example, I have discussed how the terms transparency and explainability are often used interchangeably in the academic literature in §2.1.5.

Building upon Cobbe et al.’s definition of meaningful transparency [68, 252], I argue transparency becomes *meaningful* to clinicians when the system or its designers provide, according to the clinician’s needs, information on data, goals, outcomes, compliance, influence, usage and the algorithms employed [199, 108, 364, 191], and this information is *contextually appropriate* for them. As discussed, transparency can be detrimental in some circumstances [5, 196, 359], e.g. it can sometimes clash with privacy, which is crucial to protect medical data [211, 122, 359]. It is therefore important to understand what constitutes meaningful transparency for clinicians, so as to make sure we build AI systems that meet their information needs. I now ground this investigation in a specific clinical domain: gynaecological imagery.

5.2.2 AI in gynaecological imagery: an under-researched application domain in the XAI literature

Among all clinical domains, I argue it is urgent for the FAccT, XAI, and HCI communities to consider the transparency needs of gynaecologists. It is beyond the scope of this study to review all current and potential applications of AI in gynaecology, so I focus here on the use of AI in gynaecological imagery, such as ultrasound (US) and Computed tomography (CT) scans, which are ubiquitous in the field [100] (see also §5.5.2). Combining X-rays and tomographic reconstruction algorithms, CT scans are an example of algorithmic systems introduced into medical practice the 1970s, mostly for diagnosis [73]. US scans are widespread in gynaecology as well [183], yet a 2020 literature review summarises: “AI has had a little impact on this field so far” [100].

Despite the lack of substantial impact to date, AI research for gynaecological imagery has been growing in the last 5 years, e.g. on magnetic resonance imagery (MRI) [162, 242], CT [48], and US scans [223, 148, 229]. A 2023 literature review found 41 papers using AI on gynaecological US scans, with some new sub-fields emerging from 2019 [183]. There has also been some research towards integrating such systems into clinical practice [118, 14]. Despite this recent interest, and the rise of XAI research in healthcare more generally [244, 17], few XAI publications focus on gynaecological imagery as an application domain: a 2023 systematic literature review mentions only one paper in the field, which uses convolutional neural networks to achieve an explainable diagnosis method of cervical cancer using pathological images [244, 340]. Gynaecologists’ transparency needs thus deserve further attention. I aim to contribute towards better understanding these with a

case study and co-design approach. The next section considers recent co-design and HCXAI research with clinicians from other fields to guide my experiment with gynaecologists.

5.2.3 HCXAI literature and co-design with clinicians

Human-centred AI (HCAI) has increasingly been studied by HCI and AI researchers in recent years [286, 60, 2]. I build on this HCXAI approach, as described in Chapter 2. The question of who AI is made for has been the focus of a growing number of HCI research that explore clinicians’ perception of AI systems [57, 370, 315, 376]. For example, Bussone et al. [53] explore the role of explanations on misplaced trust and over-reliance in Clinical Decision Support Systems (CDSS). Recently, Verma et al. interviewed medical-imaging experts to “scrutinize physicians’ engagement with AI (...) and disentangle its future alignment across the clinical and research workflows, diverging from the existing ‘one-size-fits-all’ paradigm within Human-Centered AI discourses” [350]. Whilst such papers inform us on how to build clinical AI systems, defining transparency meaningful to clinicians remains unanswered [336, 385, 31].

To further account for AI users’ perspectives, co-design studies have been shown to be a viable approach [241, 14], as outlined in Chapter 2. Moreover, co-design methods are widely used in health [316]. Among other, co-design have been used to build various (explainable) clinical AI systems [198, 371, 355], e.g. a machine learning-based predictive CDSSs to analyze health records [306], and a trustworthy deep learning-based skin lesion classifier [386]. These efforts include human-centred approaches: for instance, Panigutti et al. recently co-designed an explainable AI technique and user interface for CDSSs with medical professionals [271]. While others have focused on defining and addressing the challenges hindering AI implementation in healthcare [14, 188], I build on these studies and the design recommendations offered in the recent HCXAI literature as a starting point to explore what constitutes *meaningful* transparency for clinicians, towards bridging the gap between research and design implementation. I now outline these recommendations and my co-design methodology.

5.3 Methods: Co-design workshop and interviews with gynaecologists

To bridge the gap between research and the implementation of clinical AI systems, I explore clinicians’ understanding of meaningful transparency by testing and expanding recent HCXAI recommendations. I ground my investigation by using the real-life, prospective case study of a Deep Learning model to help diagnose ovarian cancer (§5.3.2) with n=15 gynaecologists. I first detail the HCXAI literature reviews I based it on, before describing

my case study and co-design activities, i.e. an in-person workshop (§5.3.3), and individual interviews (§5.3.4). This study took place in the UK and the Netherlands from July until August 2023.

5.3.1 The HCXAI literature reviews tested and extended in the co-design activities.

I start my investigation by analysing the recent HCXAI literature and using it as a broad database of design recommendations currently available for building transparent clinical AI for users. By doing so, I aim to help operationalise such recommendations and pave the way for more responsible AI in healthcare. Because HCXAI research is wide and spans various disciplines, including social sciences [232], I base the co-design activities on recent systematic literature reviews relevant to the case study (§5.3.2) and research question. I do not claim to perform a meta-analysis of such literature, nor do I have the space to discuss each HCXAI literature survey in detail. Instead, I compiled a total of 28 design recommendations for users from the HCXAI literature in three steps. First, I identified 10 relevant literature reviews published in the last four years [197, 355, 356, 10, 373, 312, 244, 152, 298, 17]. Based on i) their recent publication date (2019-2022), and ii) them explicitly offering design recommendations to build transparent AI systems for users, we selected among these three reviews [197, 355, 356]. In doing so, I excluded relevant reviews where design recommendations were not clearly provided, such as [17]. Lastly, I extracted all the recommendations from these three reviews, as reported in Tables 5.1, and 5.2, and tested them all in the card sorting activity.

Laato et al.’s “recommendations for end-users” [197] do not specifically focus on a clinical context, while Wang et al. (2019) [355] and Wang et al. (2023) [356] do. Moreover, Laato et al [197] cite both Wang et al. (2019) [355] and Xie et al. [370] in their systematic literature review, and yet I identify respectively six and two design recommendations in these publications that have not been included in Laato’s list. I thus include Wang et al. (2019)’s recommendations in Table 5.2. Although I identified two design recommendations from Xie et al. [370], I excluded them from the study for it is not a systematic literature review, and my aim is not to make an exhaustive list of design recommendations but rather provide an informative one. Moreover, please note the minor edits I used in the list of recommendations above to be able to test it with gynaecologists: i) I use British spelling throughout for consistency and removed explicit mentions of specific application domains, not to bias participants; ii) Wang et al. (2019)’s recommendations (nb. 17-22) have been shortened for clarity and usability, e.g. I removed examples and references; and iii) Wang et al. (2023)’s recommendations (nb. 23-28) were extracted from paragraphs of text by keeping the title (as reported in the Recommendation column) and one or two key sentences

Table 5.1: List of the 16 design recommendations for users provided in Laato et al. [197] (nb. 1-16)

Recommendation	Reasoning
1. Context is everything – There is no one-size-fits-all type of solution	What to explain is dependent on several factors including what kind of AI system or decision we are explaining, who are the target audience and do we want to optimise for trust, for understandability or do we wish to simply comply by legislation
2. Provide explanations on demand, not all the time	For certain decisions and in certain moments users' may be interested in seeing more information on AI system decisions. However, constant display of full XAI documentation can hurt the user experience
3. Personalise explanations	There are various kinds of people with different levels of understanding of AI systems and XAI needs. This could be taken into account when explaining the system
4. Consider visualising explanations	Users tend to anthropomorphise AI and may benefit from human-like explanations. Visualising explanations may help some users to accept the AI system and its decisions better
5. Acknowledge the existence of trade-offs	For example, optimising explanations for understandability can lead to less details, which can hurt end users' confidence in the explanation
6. Consider potential misconceptions	Users may end up forming or having formed misconceptions regarding the AI system. These may shape behaviour and interpretation of explanations in a certain way. Explanations that are able to reshape misconceptions in a constructive way of conceptual change are valuable
7. Link explanations to users' mental models	This makes the AI system easier to understand for end users, increasing transparency
8. Strengthen users' curiosity towards the system	To increase user satisfaction, provide interesting and even surprising elements to keep the users' curiosity at a high level
9. Ensure the visibility and discoverability of explanations	Make sure AI system end users find and become aware of explanations
10. Use metaphors to demystify how AI systems work	Metaphors can be more useful in increasing end users' understanding of AI systems than precise but difficult technical language
11. Support users' own thinking	The AI system should provide counterfactuals and explanations so users can reflect on and test their own thinking and hypotheses
12. Provide access to source data	Users may want to request access to raw data to build their trust in the AI system
13. Provide users with generalised explanations rather than case-based explanations	Users may consider it quirky if the decision is explained to them with a particular event from the past. To increase user acceptance, refer to generalised past events instead
14. Consider what part of the AI system to explain	Depending on the situation, users may wish to know more about, for example: (1) inputs; (2) outputs; (3) application; (4) situation; (5) model; (6) certainty; and (7) control
15. Explain unfavourable decisions	Users are likely to demand explanations when they disagree with the system
16. Communicate the uncertainties involved in the system's decision making	If there is a mismatch between users' expectations of the AI system and its actual capabilities, it hinders users' acceptance and trust building in it. Users should understand the risks of the AI system's making errors

(the Reasoning column), while removing academic references and medical connotations, so as not to bias the participants. Consequently, not all these recommendations emerged from user studies in a clinical field, but I kept the full list to enable gynaecologist participants to consider how they would categorise them.

5.3.2 Case study: Deep Learning segmentation of ovarian cancer in CT scans

To ground the co-design workshop and interviews in gynaecologists' clinical practice, I use an open-source Deep learning model built on CT scans that segments multi-site lesions

Table 5.2: List of the 12 design recommendations for users provided in Wang et al. [355] (nb. 17-22) and Wang et al. [356] (nb. 23-28)

Recommendation	Reasoning
17. Support hypothesis generation	The system should provide explanations to allow users to generate and test hypotheses to further narrow down potential causes
18. Support forward (data-driven) reasoning	Avoid confirmation bias and backward reasoning, where the user does not consider other hypotheses independently
19. Support coherent factors	Users may expect some features to be correlated or have some other relationship and would be confused if these features contradict their typical relationship. Such feature attributions should be aggregated together or have their interaction relationship visualised
20. Supporting access to source and situational data	When adopting a new AI, users may want to manually perform decision making with a few instances to build up their trust. Showing raw data or supplementary data about the aggregated together or have their interaction situation, even if not used directly by the AI, can help with this verification goal
21. Support Bayesian reasoning	It is important to show probabilities, namely, (1) prior probability to indicate the prevalence of classes in general, and (2) intermediate posterior probabilities, where after filtering on a set of salient features or factors to indicate the conditional prevalence of an outcome
22. Integrating multiple explanations	Users employ a diverse range of XAI facilities, to reason variedly. Therefore, more work is needed to integrating multiple explanations into single explanations
23. Provide additional information about AI	For example, the importance of providing a holistic, global view of the AI to users during the onboarding process, such as the system's capabilities, functionality and design objective
24. Align AI design with current workflow	Design AI in a way that can seamlessly fit into the local context and workflow. For example, making the AI "unremarkable"; that is, embedding the system into the point of decision-making and the infrastructure of the organisation in an unobtrusive way
25. Consider social, organizational, and environmental factors	For instance, it is suggested to build accountable relationships with the leadership, engage stakeholders early and often, rigorously define the problem in context, make the system an enterprise-level solution by involving other departments and units (e.g., security and legal departments), and create an efficient communication mechanism with end users
26. Respect professional autonomy	Allow users to freely operate their professional judgment and decision-making without any interference
27. Adopt a human-centered design approach	Create ongoing feedback loops with users and stakeholders. By doing so, researchers and system designers can continuously and iteratively collect user feedback to inform the design of AI
28. Other considerations	Providing in-depth training to users to flatten the the learning curve, designing AI as a multi-user system to better engage users in decision-making, expanding the application of AI tools to a variety of scenarios, and examining the data quality for potential bias and fairness in AI

in ovarian cancer [48]. Although this model is not currently implemented in clinical pathways, I use it to illustrate the kind of models that could be potentially introduced into gynaecological imagery going forward. I used slides with a CT scan example to introduce the model and co-design activities. Moreover, most gynaecologists are familiar with ovarian cancer and CT scans, making it an easily understandable example for participants. I acknowledge cancer diagnosis is performed differently according to the country and healthcare system: in the UK for example, a team of specialised clinicians including gynaecologist oncologists, oncologists, and radiologists, is involved in discussing the CT scans. According to the country, I thus focus on the perspective of gynaecologists' and gynaecologist oncologists experienced in operating ovarian cancer patients. I now describe the co-design activities based on this case study in more detail.

5.3.3 In-person co-design workshop with gynaecologists in Amsterdam UMC, the Netherlands

I conducted a co-design workshop with gynaecologists to better understand gynaecologists' perspective about meaningful AI transparency for clinicians towards operationalising the design recommendations listed above. I recruited $n=10$ gynaecologists experienced with ovarian cancer patients in Amsterdam UMC, the Netherlands via email invitation; all knew each other, and the workshop was attended by a PhD student in XAI and facilitated by two PhD students in HCI. The workshop lasted 1h30 and after being introduced to the concept of meaningful transparency and the case study orally with slides, the participants were teamed into groups of two to three, and completed a card sorting activity in about 20 minutes. The aim of the card sorting activity was to evaluate to what extent the design recommendations currently offered in the HCXAI literature encapsulates the concerns of clinicians, and explore the sorting and category labels participants identify based on these recommendations. It included 28 cards directly copying the design recommendations from the HCXAI literature reviews described in §5.3.1. Each card was printed on a A4 paper sheet with a title and a short description on the reverse. Each team sorted the cards into categories of their choice, which they named without prompts. These were discussed, challenged, and sometimes merged in the second activity.

The second activity consisted of co-designing a 'transparency map': I copied all the categories identified by participants in the card sorting activity onto large sticky notes and co-designed a conceptual diagram with the whole group, which I called a map for simplicity. I placed the sticky notes on a wall so that physical proximity between the notes represented the conceptual closeness of the categories, e.g. superimposing identical categories. Participants collectively decided where to place the categories on the map. When consensus did not emerge quickly, I duplicated the category to reconcile disagreements on their location. Once all the categories were placed on the wall, the participants were given the opportunity to add, remove categories, and edit the map. The aim of the transparency map is to i) challenge, edit, and validate the main categories identified by participants in the card sorting activity (based on HCXAI design recommendations for building transparent AI systems for users); ii) highlight the connections they saw between these categories as indicative of their understanding of meaningful transparency; and iii) visually represent gynaecologists' understanding of meaningful transparency to extract insights that might lead towards practical design implementation. This activity was audio recorded and transcribed by one author. Finally, the workshop participants filled a short evaluation form focusing on their lived experiences of the co-design activities. I also asked about their experience with clinical AI, attitude towards AI in general, as well as demographic information.

5.3.4 In-person interviews with gynaecologists in the UK and the Netherlands

To confirm the data collected during the workshop, I conducted $n=6$ in-person, individual interviews with gynaecologists. I recruited participants via email invitation in hospitals in the Netherlands and the UK. As a result, only one participant took part in the workshop just before their interview in Amsterdam UMC with two PhD students in HCI. All remaining interviews were conducted in the UK by one PhD student in HCI with gynaecologists who had not attended the workshop. I recruited mostly ($n=4$) gynaecologist oncologists to reflect the sub-specialisation of clinical practice in the UK and the specific transparency needs of this user group, though all $n=6$ participants had experience with ovarian cancer patients. These took place in the Nuffield Health Cambridge Hospital and Addenbrooke's Hospital in Cambridge, Queen Elizabeth Hospital in London, and Norfolk and Norwich University Hospital in Norwich, UK.

The interviews lasted 45 minutes and consisted of one short training on decision-making and AI with slides and two co-design activities. After being introduced to the definition of meaningful transparency and the ovarian cancer case study (for the five interviews in the UK only), I discussed the theory of system 1, or intuitive, fast thinking process, and system 2, or analytical, slow thinking process [115, 184, 331, 355] (for all $n=6$ interviews). 'System thinking' theory is largely considered relevant in user research with clinicians [331, 355]. It was thus chosen to set context as an initial prompt to help participants create 1-2 gynaecologist *personas*: characters who represent different user types that they had encountered in their professional environments. They were asked to briefly describe the persona's usual thinking style during the diagnosis process, whilst thinking about the case study. They could describe them referring to system 1 and 2 if helpful. For this activity, I printed forms titled 'Person description.' The interviewer(s) created 1-2 personas as well, and reproduced the workshop's transparency map by sticking on a wall large fresh sticky notes (so they would not fall during the activity). This persona activity helped investigate decision-making processes used by gynaecologists, and generate different clinical AI users based on participants' real-world experience in healthcare settings.

After discussing their personas together, the participant and interviewer(s) critiqued and co-edited the map in relation to the personas identified; the 'person description' forms were placed on the map above the category (i.e. the group of HCXAI design recommendations) that seemed most relevant for this persona according to its creator. This activity allowed to confirm previous results and elicited further discussion. Finally, the interviews concluded with a debrief and the same evaluation form as the workshop's. The interviews were audio recorded and transcribed by one author.

5.3.5 Gynaecologists’ demographics

Table 5.3 outlines the total n=15 participants’ demographic information. Three participants (P13, P14 and P15) took part in the workshop (and were included in my analysis) but did not provide demographic information. The participants comprised six consultant gynaecologists, four consultant gynaecologist oncologists, two resident gynaecologists, and three other gynaecologists (P13-15), all 15 having surgical experience. This sample takes into account the sub-specialisation and targets in place in the two countries, e.g. NHS patients with suspicion of ovarian cancer should be referred to an gynaecologist oncologist within two weeks. Most participants declared having a strong interest but limited experience and understanding of AI systems. Most state they use no AI system in their work – though one specific clinical AI system called IOTA is mentioned in Amsterdam. Finally, most declare ignoring how AI systems work, or describe them as: “magic,” and “no human errors.”

Table 5.3: Participants’ self-described roles, gender, years of experience, co-design participation, and country. Participants are referred as P[number] in the following sections.

ID	Self-Described Role	Gender	Experience (years)	Participation	Country
P1	consultant gynaecologist	Woman	11	Workshop	Netherlands
P2	consultant gynaecologist	Woman	Not given	Workshop	Netherlands
P3	consultant gynaecologist	Woman	20	Workshop	Netherlands
P4	resident gynaecologist	Man	5	Workshop	Netherlands
P5	consultant gynaecologist	Woman	Not given	Workshop	Netherlands
P6	resident gynaecologist	Woman	Not given	Workshop	Netherlands
P7	consultant gynaecologist	Man	over 15	Both	Netherlands
P8	consultant gynaecologist oncologist	Man	13	Interview	UK
P9	consultant gynaecologist	Man	17	Interview	UK
P10	consultant gynaecologist oncologist	Man	40	Interview	UK
P11	consultant gynaecologist oncologist	Man	15	Interview	UK
P12	consultant gynaecologist oncologist	Man	8	Interview	UK
P13	gynaecologist	Not given	Not given	Workshop	Netherlands
P14	gynaecologist	Not given	Not given	Workshop	Netherlands
P15	gynaecologist	Not given	Not given	Workshop	Netherlands

5.3.6 Ethical considerations

For this study, I did not collect any medical data and urged all participants not to mention any information regarding their patients throughout the study. Additionally, I design the co-design activities keeping in mind the group dynamics and potential hierarchy in place within a team of clinicians. I discussed this challenge with other clinicians from my network and chose to conduct the group workshop in the Netherlands for the healthcare environment has been described to me as less hierarchical than in the UK. In practice I also designed the final survey to make sure all participants were able to rate their co-design experience and add further comments and feedback should they wish to. No further ethical

risk was identified for participants joining this study.

5.3.7 Data analysis

The co-design activities produced qualitative and quantitative data. The card sorting activity was analysed by counting the cards, categories, and comparing the labels and groupings made by the four teams. I compared the outputs to the three papers [197, 355, 356] the cards were copied from. The transparency maps produced during the workshop and interviewees were compared to one another and in relation to the card sorting results and HCXAI literature reviews. I used Clarke and Braun’s thematic analysis method [45] to extract key themes from the audio recordings and transcripts. I conducted all six stages (i.e. familiarizing yourself with the data, generating initial codes, searching for themes, reviewing themes, defining and naming themes, producing the report) over two iterations. I searched for semantic themes in the transcripts and audio recordings using an inductive, data-driven approach. Together with another co-author, we discussed, and validated the themes, before repeating the process. I kept participants’ wording, as reported in Table 5.4. Finally, the demographic data collected, the information about participants’ experience with AI, and their experience of the co-design activities helped contextualise the findings. I now present my main results.

5.4 Results

I ran a co-design workshop and interviews with a total of n=15 gynaecologists towards operationalising current HCXAI recommendations in healthcare. I now describe the themes they raise and key insights into their understanding of meaningful transparency for clinicians in the context of the ovarian cancer case study.

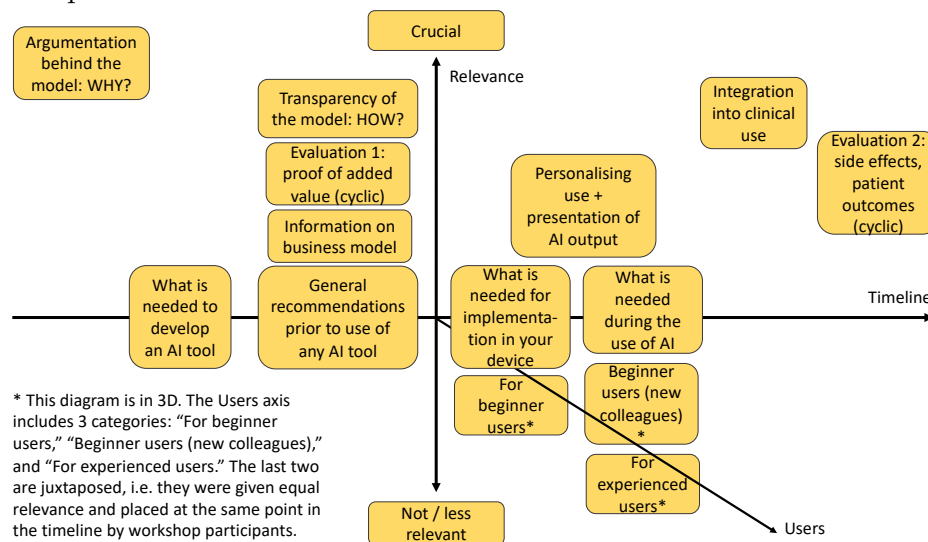
Table 5.4: Summary of main themes raised by participants (in random order).

Overarching Category	Themes
Timeline	Before/during/after implementation, integration into clinical practice, during use, pre-op planning & during surgery
Importance	More/less crucial, “the why [using AI]” more important than “the how [AI works]”
Users	Experienced & beginner [tool] users, specialist & generalist gynaecologists, trainees & senior doctors, fast- & slow-thinkers, personality traits & cultures, expectations & medical norms, decision-making & communication styles, approaches to surgery, multidisciplinary team (MDT)
Cyclic process	Evaluation (e.g risks for patients, patient outcomes, trials), feedback loop (e.g. updating the data), collaboration with clinicians, cancer-specific diagnostic process
User interface	Communication with users, clinician-friendly & intuitive, personalisation & levels of information, risks (i.e. biased data, cumbersome information), a key to look into the system, interactive tool
Challenges	Unusable systems, trust in system/model, trustworthy colleagues, tech skepticism, national & regional IT disparities
Questions on transparency	difference between transparency & explanation, relationship between transparency & usability

5.4.1 Main themes raised by participants

Table 5.4 summarises the main themes raised by participants during the workshop and interviews. I group these into seven categories, named with participants' wording: 'users,' 'relevance,' 'timeline,' 'cyclic process,' 'user interface,' 'challenges,' and 'questions on transparency.' There was a consensus among workshop participants and interviewees regarding the importance of the first three axes, which therefore became the focus of this study. These three transparency categories are visualised in Fig. 5.1, showing (i) the distinction between the information needs of different user groups among gynaecologists (Users axis), (ii) the relevance of various types of information (Relevance axis), and (iii) the chronological order of the information users receive (Timeline axis).

Figure 5.1: Diagram representing the transparency map co-designed with gynaecologists in the workshop in Amsterdam.



5.4.2 Axis one (Users): differentiating the information needs of different gynaecologist user groups

Participants stressed the importance of differentiating the information needs of various types of gynaecologist users among clinicians likely to use an AI-based tool to help diagnose ovarian cancer. Differences in medical expertise, clinical experience, decision-making processes, and habit of using a given AI tool were highlighted. I argue that a level of granularity in analysing and considering users is needed to add nuance to the design recommendations described in Tables 5.1 and 5.2 and to help operationalise them. For example, interviewees raised these distinctions when creating gynaecologist personas: out of the 18 identified, 8 were described as using mainly system 1, i.e. a fast and intuitive thinking process, 7 as primarily relying on system 2, i.e. a slow, analytical thinking

process, and 3 as using a combination of both [184, 115]. An interviewee notes: “I think a chronological thinking process is very apparent in surgeons (...) because [surgery] is a procedure (...) I think if you do any intervention, you will develop a [thinking] structure in steps” P11. These were often paired with personality traits, e.g. quiet and thorough, or social and dogmatic. By placing their personas on the transparency map, participants revealed how different gynaecologists can prioritise different information needs, in their experience (see §5.4.4). Another interviewee commented: “we [gynaecologist oncologists] don’t need the same level of information as generalist [gynaecologists]. To me, [information about] the risk [for a patient] is more crucial” P8. Similarly, workshop participants drew a clear distinction between “beginner users” (including “new colleagues”) and “experienced users” in terms of interaction with an AI tool (Fig. 5.1), and all interviewees confirmed it. These different users often work closely in teams, as one interviewee summarises it: “each individual [in an Multidisciplinary team (MDT)] has a specific set of expertise and, jointly, it’s like the pieces of a puzzle coming together” P12. The HCXAI recommendations from prior work (Tables 5.1, 5.2) do not distinguish these various types of users, nor reflect their collaborative interactions with technology. Connecting them to specific types of clinicians can help operationalising them [350]. By providing leads to define such users, I thus contribute towards facilitating the implementation of more clinical AI systems.

5.4.3 Axis two (Relevance): providing information in a prioritised manner in terms of perceived relevance for clinicians

As visualised in the second axis (Fig. 5.1), participants expressed the need to receive information about a clinical AI tool in a prioritised manner in terms of perceived relevance for them. Importantly, their priorities seem to differ from some of the recommendations from recent HCXAI literature. For example, interviewees placed virtually all the personas on the categories higher up on the map, which confirms their perceived importance for gynaecologist users. Indeed, the three main axes of the map reveal that the timeliness and relevance of the information provided about an AI-based tool throughout the AI tool’s design lifecycle is key in participants’ understanding of meaningful transparency for clinicians. This prioritised and user-centric logic has not been described nor visualised for this context in the HCXAI literature yet. This insight is key to make sure we design AI systems that are meaningfully transparent for clinicians going forward. Moreover, when workshop participants grouped the 28 design recommendations extracted from the HCXAI reviews (Tables 5.1 and 5.2) into the category of their choice (unprompted), some recommendations were grouped as “not that relevant.” Two teams (50% of workshop participants) grouped 11 recommendations (i.e. over a third of the recommendations) as “not/less relevant” and four of these recommendations were mentioned by both teams as

being of lower relevancy: “Personalise explanations” (nb. 3), “Use metaphors to demystify how AI systems work” (nb.10), “Provide users with generalised explanations rather than case-based explanations” (nb. 13), and “Integrating multiple explanations” (nb. 22). While the recommendations nb. 3, 10, and 13 are not specifically listed for clinical users in the HCXAI literature reviews (Tables 5.1, 5.2), “Integrating multiple explanations” is [355]. Participants therefore challenge these recommendations for clinicians. Finally, the transparency map reveals another gap between the HCXAI literature and gynaecologists’ perspective on meaningful transparency. Workshop participants seem relatively less concerned by the “what to explain” and “how to explain” categories mentioned in Laato et al. [197] than the “why,” and “integration into clinical practice” categories they identify during the card sorting activity, as well as the additional “evaluation” categories they introduce during the map co-design activity (§5.4.4). Interviewees equally emphasise these categories, but none of the three selected literature reviews highlight them for clinicians [197, 355, 356]. This shows that, whilst recent HCXAI literature provides a starting point to discuss meaningful transparency for clinicians with gynaecologists, there are differences between some HCXAI recommendations and what gynaecologists consider as most relevant to them, and thus contextually appropriate transparency for clinicians.

5.4.4 Axis three (Timeline): prioritising information about the AI-based tool’s conceptualisation, role, and purpose over (post-hoc) explanations

The third axis, called ‘Timeline’ in the transparency map, represents the stages in the AI design lifecycle when gynaecologist participants want to receive transparent information about the AI-based diagnostic tool (Fig. 5.1). Combined with the second axis (Relevance), it reveals that participants find information provided to them *before* the AI tool gets implemented more important for gynaecologists than (post-hoc) explanations, such as how and why it reaches a certain output with a given input. The former type of information includes among others insights about why the tool can be helpful to diagnose a given patient, what help they can expect from it, and why they should use it. Indeed, the category “Argumentation behind the model: WHY?” was placed by workshop participants before and higher than every other category on the transparency map, including above “Transparency of the model: HOW?” (Fig. 5.1). One workshop participant explained: “you need to have an idea why you’re going to start this entire project” P3. Moreover, whilst all interviewees independently selected the “evaluation,” and “how?” categories as most relevant for personas described with system 2 (slow-thinking), and “integration into clinical practice” and “personalising use” for personas described with system 1 (fast-thinking), “the Why?” was once again considered by interviewees the most important category of

transparent information when considering all types of personas (described with system 1, system 2 or both). Gynaecologists’ preference for such type of transparent information might be partly explained by one interviewee’s comment on the information needs of beginner and experienced users, who mainly use system 1 thinking: “I can imagine an interface that in the beginning gives me more information and in the end it ends up in a small question mark at the bottom. And if I use the model ten times, it’s going to decrease the amount of background information that it’s giving me, and it’s more and more just showing me the results (...) one of my colleagues, they always want the background information. I don’t, if I trust the system, I’m fine” P7. This confirms what Burgess et al. recently call “the front-loading” trust, whereby clinicians want “to determine their trust of an AI insight system when first introduced to the tool” [50]. However, the emphasis on holistic transparent information provided prior to the AI tool’s implementation into clinical practice as a mechanism to support such trust was not in the design recommendations from the recent HCXAI literature reviews (Tables 5.1, 5.2). It warrants further research to be confirmed, and seems to indicate we need to rethink how and *when* to provide transparent information to such clinicians.

5.5 Discussion

After testing the recommendations currently offered in the HCXAI literature (Tables 5.1, 5.2) and providing insights into gynaecologists’ understanding of meaningful transparency (Table 5.4) in the context of the ovarian cancer case study (§5.3.2), I suggest four design recommendations for meaningful transparency for clinicians (§5.5.1). I then discuss the importance of direct engagement with gynaecologist users in relation to women’s health and HCI and AI research (§5.5.2), before outlining the study limitations (§5.5.3) and research avenues to bridge the gap between HCXAI literature and implementation of transparent AI systems for clinicians (§5.5.4).

5.5.1 Design recommendations for meaningfully transparent AI systems for clinicians

Based on the findings, Table 5.5 outlines my four design recommendations for transparency meaningful to clinicians.

5.5.1.1 Prioritising information categories about the why, evaluation, personalisation, and integration into clinical practice.

Given what participants expressed as their main information needs (Fig. 5.1), my first design recommendation moves beyond the list of 28 recommendations I have extracted

Table 5.5: Four design recommendations for AI systems meaningfully transparent for clinicians (in no particular order).

Design Recommendation	Reasoning related to my case study with gynaecologists
1. Prioritising information categories about the why, evaluation, personalisation, and integration into clinical practice	This focus can help cover the information needs of most personas I have identified with gynaecologist participants (§5.4.2, §5.4.3, §5.4.4)
2. Providing prioritised information in terms of timeliness and relevance for clinicians throughout the AI design lifecycle	This must align with gynaecologists’ thinking process (Fig. 5.1, §5.4.3) and clinical experience as surgeons
3. Accounting for multidisciplinary teams (MDT) using a clinical AI system by taking into account the perspective, role, and human interactions of various types of gynaecologists	For example, in my case study, by better defining the role of gynaecologist oncologists in discussing with radiologists the CT scans towards reaching a pre-operative plan (Table 5.4, §5.4.2)
4. Providing transparent information on the AI-based tool’s conceptualisation, role, and purpose prior to its implementation	Information should be provided in particular before the tool gets implemented into clinical practice, e.g. during the initial clinical trials (Table 5.4, §5.4.4)

from the HCXAI literature by prioritising four categories of transparent information: “the why,” “evaluation,” “personalisation,” and “integration into clinical practice.” Participants created the ‘evaluation’ category during the map co-design activity, so it is not directly connected to some pre-existing HCXAI recommendation listed in Tables 5.1 and 5.2. It includes, among others, information about the risks for patients, the patient outcomes, and the cost effectiveness of the tool (Table 5.4). This category is key to better align HCXAI design recommendations with real-life clinical practice, and is likely applicable to other clinical domains as well. However, gynaecologist participants did not mention the need to embody ethics and liability nor to develop AI educational opportunities, unlike Verma et al.’s oncologist interviewees [349]. This might reveal another specificity about gynaecologist users, who may be less exposed to clinical AI than those in other specialities.

5.5.1.2 Providing prioritised information in terms of timeliness and relevance for clinicians throughout the AI design lifecycle.

My second recommendation aims to align transparency with clinicians’ and, in particular, surgeons’ structured thinking process and clinical workflow (see §5.4.2). This confirms previous research on surgeons’ procedural approach to tasks [151] but also AI onboarding [57]. Note my participants are surgeons, but not all gynaecologists are. I argue transparency can become more meaningful to such clinicians by following a structured order, key dimensions to which are visualised in Fig. 5.1, though further research could assess whether this also applies to other types of gynaecologists (Table 5.4).

5.5.1.3 Accounting for multidisciplinary teams (MDT) using a clinical AI system by taking into account the perspective, role, and human interactions of various types of gynaecologists.

My third design recommendation specifies the HCXAI emphasis on accounting for the socio-technical context when designing and evaluating an AI-based system for users [111, 108, 68] in relation to a less-studied population and clinical domain—gynaecology (§5.2.2)—and uses a real-life case study to ground my findings into real-world settings [14]. Indeed, in the UK, gynaecologist oncologists approach ovarian cancer diagnosis and treatment in MDTs, and thus an AI-based diagnostic tool needs to support their interactions within such teams (see §5.4.2). By describing various gynaecologist users of a potential, real-life AI-based CDSS (§5.4.2), I follow Verma et al. in departing from the “one-size-fits-all” paradigm within HCAI research [350], and Berg in arguing that clinical IT development requires a user-centred approach, due to the complex network of people and practices in healthcare [30]. Further research is needed on characterising the transparency needs of gynaecologists as individual clinicians and within MDTs. Others have also stressed the need to better support patient-provider collaboration [178, 50].

5.5.1.4 Providing transparent information on the AI-based tool’s conceptualisation, role, and purpose prior to its implementation.

My final recommendation stresses not only the need for designers to communicate and engage with clinical users iteratively throughout the design process [30], but also the importance of providing them with transparent information about the AI-based tool’s conceptualisation, role, and purpose *before* its implementation (§5.4.4). This could perhaps help to keep AI-based CDSS less obstructive and increase their contextual fit [374]. It also suggests that, in order to improve transparency for such clinicians, HCXAI research should shift from focusing predominately on XAI to also exploring more inclusive approaches, and provide clinicians with information before an AI tool gets implemented (and, indeed, throughout the entire AI lifecycle, so as to enable understanding and scrutiny [68]).

5.5.2 Women’s health in the HCI and AI literature: engaging directly with gynaecologist users

One key contribution of this study is to expand the AI and HCI literature on women’s health by directly engaging with gynaecologist (professional) users. In doing so, I build upon the growing body of literature exploring women’s bodily experiences and interactions with health and wellbeing technologies [11, 189], such as Homewood et al.’s innovative and phenomenological period tracking designs – e.g. Ovum [160, 161] and Ambient Cycle [159]– and Bardzell’s feminist HCI research [157, 26, 27]. As I explicitly focus

on gynaecologists’ information needs in the context of a real-life case study of ovarian cancer, my study can thus complement such feminist HCI approach by contributing to implement more transparent (and thus responsible) clinical AI to advance women’s health. Indeed, I aim to promote further HCXAI research relevant for improving women’s health realities, where both the medical practitioners and patients’ perspectives are critical for realising the potential of clinical AI. Moreover, and in line with the limitations of AI transparency (§5.2.1), I also acknowledge that clinical AI can bring some risks for women, e.g. discrimination. Thus more HCI and AI research is needed into “inaction as a design decision” [158] in relation to women’s health, i.e. where designers decide not to implement AI because of its risks for users.

5.5.3 Study limitations

Due to the logistical challenges of organising in-person, co-design activities with gynaecologists, one limitation of this study is the number of participants I was able to recruit. There is also a self-selecting bias in my sample of participants, based on their interest in AI systems. However, co-design can yield insights when conducted with a small group of participants, and we reached consensus on most of the activity findings. I also acknowledge that despite my best efforts to reproduce each experiment in the same conditions, this was not always possible. For example, the map co-designed during the workshop was reproduced in each interview on a different wall with different features. However, each interviewee had equal opportunities to discuss, challenge, and critique this map. Similarly, there was insufficient time with clinicians to discuss system 1 and 2 thinking in the workshop, but this was fully covered in the interviews. Moreover, I based the co-design activities, e.g. card sorting, on three recent HCXAI systematic reviews that explicitly included design recommendations for users [197, 355, 356]. Thus, I do not claim the list of 28 recommendations based on these surveys is exhaustive, e.g. I identified two additional ones in Xie et al. that were not tested here [370]. I argue this list is nonetheless indicative of the types of recommendations currently offered in the HCXAI literature to build transparent AI systems for users. Indeed, explainability and transparency are different concepts, but given the various use of both terms in the academic literature, I used the HCXAI literature reviews as broadly indicative of the current research done in algorithmic system transparency. In all, my findings towards operationalising such recommendations can be confirmed with more studies in the field, as described below.

5.5.4 Further research towards implementing meaningfully transparent AI systems for clinicians

Lastly, I encourage more co-design research into implementing meaningfully transparent AI systems for clinicians. Indeed, Thieme et al. show it can effectively lead to the integration of a production interface [334]. Patients must also be engaged as critical stakeholders of clinical AI. This was outside the scope of the study, however I encourage more studies directly engaging with patients of clinical AI, in particular women and non-binary individuals, to help close the gender data gap [55, 54, 276]. I suggest for example further studies to validate my design recommendations for meaningfully transparent clinical AI systems with a different gynaecological case study (§5.5.1). As participants have unanimously highlighted the importance of the timeline when designing such clinical AI system, I also encourage future work to involve gynaecologists as co-designers as early as possible, and at every stage of the design process. Indeed, a 2021 systematic literature reviews shows that “clinical experts are less prevalent in developmental stages to verify clinical correctness, select model features, preprocess data, or serve as a gold standard”[306]. Building upon my findings, engaging clinicians as co-designers might be particularly relevant to design for the “why,” “personalisation,” “integration into clinical practice,” and “evaluation” categories described above (§5.5.1). Moreover, I have described and visualised (Fig. 5.1) how some gynaecologists understand what constitutes meaningful transparency in clinical AI, in the context of ovarian cancer diagnosis with CT scans (§5.4.3). However, as this study has shown (§5.4.2), there is a need for future research to account for more detailed descriptions of clinical AI users, for example by focusing on the perspective and information needs of clinicians who might not be as interested in AI systems as my participants. Finally, although I have explored how transparency could become meaningful to clinicians, investigating contexts when algorithmic system transparency might not be enough or could fail, for example to facilitate accountability or calibrate trust, was beyond the scope of this study. This question is equally important for safe and ethical AI application in healthcare, and requires further research with clinicians, including various types of gynaecologists.

5.6 Conclusion and main takeaways

In this chapter, I have provided empirical insights towards addressing the Guideline Challenge (RQ2), i.e. recent guidelines and design recommendations must align with designers’ perspective and users’ needs (Table 1.1). To better operationalise the design recommendations from the recent HCXAI literature (Literature gap 3), I have investigated gynaecologists’ understanding of *meaningfully* transparency for clinicians with a real-

life case study (Literature gap 1). Indeed, to this day, implementing AI research into clinical practice remains challenging, and rather limited in gynaecology, hence the need for more transparency in clinical AI systems. While transparency is not a panacea, it is recognised as helpful in such high-risk, real-world settings, e.g. to calibrate clinicians' trust and facilitate accountability. Towards this, I have conducted a co-design workshop and interviews with $n=15$ gynaecologists in the UK and the Netherlands. Using the case study of a Deep Learning model for ovarian cancer, I have tested and extended recent HCXAI recommendations for building transparent AI systems for users, by grounding them into a specific and less-studied clinical domain: gynaecological imagery. In doing so, I have provided empirical insights to complement and help operationalise such recommendations in clinical practice.

Overall, this chapter shows that HCXAI must better account for clinical teams with different types of gynaecologist users. I also show that the timeliness and relevance of the information given users about the AI-based tool is key for transparency to become meaningful for clinicians throughout the design lifecycle, and in particular information about the tool's conceptualisation, role, and purpose provided prior to its implementation in clinical practice. My main contributions include: i) testing recommendations from the recent HCXAI literature in a specific, relatively less-studied clinical domain with a prospective, real-life AI application; ii) describing and visualising gynaecologists' understanding of meaningful AI transparency for clinicians; iii) outlining research and four design recommendations towards meaningful transparency for clinicians; and iv) expanding HCI and AI research on women's health by directly engaging with gynaecologist as users and co-designers. Exploring such issues is key to implement more transparent AI systems that meet clinicians' information needs.

In the final study of this thesis, I return to the question –left open in Chapter 4– of how to assess the transparency dimensions of an algorithmic system (Assessment Challenge (RQ3)). I have stressed in Chapter 2 that this challenge has not been the focus of enough academic attention so far (Literature gap 4). In Chapter 3, I have started exploring this challenge by benchmarking meaningful transparency for parents in the UK by applying Contextual Integrity as a theoretical framework in relation to digital parenting. I follow this qualitative approach in the next chapter by experimenting with a critical method to assess the transparency of an algorithmic system's user interface. In doing so, I aim to better account for the broader context of users' perceptions of algorithmic systems, and thus follow recent HCXAI research discussed above.

Chapter 6

Assessing transparency: a critical method to account for the socio-technical context

This chapter is based on the following paper: Bianca G. Schor, Jatinder Singh, and Alan Blackwell. 2024. The Visual Culture of Design for Algorithmic System Transparency. An Analytic Case Study of the APB and Parcoursup Systems. In *AI & Society* (under review). I conducted all of the data collection and analysis for this project. I was assisted in the writing and editing of the paper by my PhD supervisors Dr. Jatinder Singh and Prof. Alan Blackwell.

6.1 Introduction

My final study delves further into the Assessment Challenge (RQ3) and the fourth literature gap I have identified: the lack of methods to evaluate (meaningful) transparency for users in algorithmic systems (Table 1.1). My aim is to move forward in effectively promoting more meaningfully transparent systems for users by equipping designers, policy-makers, regulators and academics with an accessible, critical, and inter-disciplinary method to assess the transparency dimensions of user interfaces. As discussed in §2.2.3.2, user interfaces are a key liminal space for non-expert users to interact with systems and often reveal gaps in transparency design of such systems. I explore this with a socio-technical approach, as discussed in §2.2.2. To do so, I adapt a well-established method from art history to the study of interface design in order to better account for the socio-technical context of transparency design. To my knowledge, this is the first attempt to introduce art history into a socio-technical approach of algorithmic system transparency. I first justify my critical approach to assess algorithmic system transparency.

6.1.1 A critical and socio-technical system approach to assess transparency

As explained in Chapter 2, there is a broad acknowledgement that transparency should be context-specific and therefore that transparency design and assessment should consider the socio-technical context of algorithmic systems. Towards this, I define them as socio-technical systems “that are composed of technical artifacts, human agents, and institutions [and]—in addition—contain artificial agents and certain technical norms that regulate interactions between artificial agents and other elements of the system” [346]. I introduce the art history methodology called object-based analysis, and argue it allows to better account for the cultural and visual background that can influence the public reception of user interface designs and their transparency dimensions. In applying it to a real-life case study, I aim to demonstrate how such analysis of visual elements can be included in a socio-technical approach to transparency for non-expert users. I also show that object-based analysis is helpful to assess a system’s transparency dimensions, based on its rigorous process that aims to make meaning of a wide range of visual rhetoric, including those composing a user interface. I argue this wide, interdisciplinary approach is a productive perspective to bring transparency theory closer to design implementation, and thus to further promote responsible algorithmic systems for users in practice.

When examining algorithmic transparency, most analysis is based mainly on textual material, whether policy documents, legal codes, or media reports [182, 171], and I argue here this is a too restrictive understanding of non-expert users’ interactions with algorithmic systems. There has been indeed some understanding that visualisation plays an important role in ‘explaining’ algorithmic systems [236, 152], although most research and design guidance (e.g. [363, 338]) focuses on the visualisation of quantitative or computational information. In this chapter, I do not consider the role of visualisation from a UX design perspective, e.g. in terms of usability [47]. Instead, I explore the potential for an *object-based analysis* of interface design based on art historical methods, i.e. a rigorous, qualitative, and holistic analysis of the visual rhetoric around transparency of user interfaces, to contribute to understanding of how algorithmic decision systems are received by society and how they mediate between actors. I argue that this method helps to better account for the influence of visual rhetoric users’ *capabilities* [307] related to transparency over algorithmic systems, which are necessary for them to make informed decision about whether and how to use such systems.

Inspired by Nobel Prize winner Amartya Sen’s work in welfare economics [307, 308], I define capabilities here as the means needed for an individual to achieve a realistic goal when interacting with an algorithmic user interface. In this context, I consider transparency and its visual rhetoric as one of the key capabilities for individuals to reach their goals, for example applying to university online (§6.3.2). My approach builds on

concerns within HCI for better integration of art-critical perspectives into the field [24], by applying a method (object-based analysis) from art history to the study of interface design. I demonstrate the potential of this approach to better account for the role of visual rhetoric around transparency through close visual analysis of two interfaces: APB and Parcoursup, the algorithmic systems for university admissions in France (Fig. 6.1).

To my knowledge, this study is the first attempt to introduce art-historical methodology into Computer Science, and the first study to focus on the transparency dimensions – as defined in Chapter 2– of the French national university admission systems. Please not that, whilst this study was completed in August 2021, Parcoursup remains a core point of contention in French public debates and the 2024 election campaigns: “The New Popular Front wants the end of Parcoursup” [16]. It thus urgently warrants more academic attention.

6.1.2 Chapter outline

In the following sections, I introduce the research background to this analytic approach and my method. I then present the case study background to Parcoursup and its predecessor APB. The main part of the chapter presents the visual analysis of these systems and of their public reception based on an established art historical method, then considers the implications of these findings for transparency. I conclude by considering the future potential of such approaches.



Figure 6.1: Welcome Pages of APB in July 2017 (left side) and Parcoursup in August 2022 (right side), on a computer screen [89]

6.2 Background: applying object-based analysis to HCI

In this research, I apply the art historical methodology of *object-based analysis* [134], which focuses on the visual rhetoric of the object of study, here user interfaces, to highlight the

potential differences between the purposes and audiences that object is supposed to serve, e.g. here increased transparency for users, and its actual design and reception. To my knowledge, this method has not been applied to non-art interfaces yet, so I experiment with it in relation to the critical inspection of system transparency. Championed by The Courtauld Institute of Art in London, object-based analysis focuses on the *visual* and/or *material* analysis of the object of study [134]. Visual analysis involves close inspection of the artwork as an image, including: its overall theme, characters and/or objects depicted; the spatial arrangement and environment; the colours, light sources, and reflections; any written or hidden symbol; the way it is framed and the perspective constructed; any artistic genre it belongs to, and so on. Material analysis, on the other hand, considers it as an object, with detailed examination of the media used (canvas, digital screens etc.); their particular components; and their role in creating the overall artwork. For example, the choice of fabric for a tapestry, can influence perceptual properties such as light reflection and contrast, as well as its monetary and socio-cultural value, which in turn enlighten us on how it was used and read by their beholders.

Object-based analysis can be comparative, iconographic, or symbolic, and complemented with archival research, literary or film criticism, or other sources of data appropriate to historical periods and research questions. These also depend on researchers' individual skills, for example the languages (modern or ancient) in which they are able to draw on sources. Comparative analysis of the features of two or more objects of study requires less professional training and is more accessible for non-art historians, than iconographic analysis for example, which requires understanding of art historical movements and genres to situate the work within art historical evolution. Similarly, symbolic analysis is often based on a religious reading of artworks [272] and therefore draws on extensive background knowledge into the values and symbols of the socio-cultural environment which the object of study belongs to. Archival research—when sources are available—can also be used to confirm interpretations through cross-referencing to build the legitimacy of the argument. In the context of this study, this offers connections to HCI scholarship [25, 38, 281].

As discussed in §2.2.3.2, I argue the user interface and its visual rhetoric is an essential object of study for assessing an algorithmic system. Although transparency in itself has limits [196, 324, 5, 13, 326], improving the transparency of systems for users through interface design is crucial as a means for identifying and in part reducing potential risks for users [69]. With this case study, the APB and Parcoursup university admissions systems, I aim to demonstrate an exemplar object-based analysis in practice. I now outline the object-based methodology as I apply it to the study of interface design.

6.3 Object-based analysis methods: accounting for the visual rhetoric of transparency

As noted, the success of a transparency regime largely depends on whether users can easily access and comprehend its constituent elements [93]. In this section, I describe, in a general sense, how to apply an object-based analysis to the study of interface design before explaining why it plays an important role in assessing how this process unfolds.

6.3.1 Object-based analysis as a methodology to study interface design

I now explain the object-based analysis of an interface as the following 8-step iterative process (Fig. 6.2) – see examples in §6.5. Please note that, to my knowledge, there is no explicit methodological guide on how to conduct an object-based analysis in art history. In my experience as an academically trained and practicing art historian, this method is often taught in art history classes and publications through examples rather than with a guide. Based on my academic training and professional experience researching various artworks, I am therefore providing here my interpretation of how to carry out such analyses in practice by identifying eight essential steps. To my knowledge, I am thus providing the first explicit guide on how to conduct such methodology in practice, and in particular in relation to a user interface:

1. Choose one or more digital object(s) of study, an audience who interacts with it, a research question, and primary visual and/or material analysis. In the case of interfaces, a visual analysis is often more relevant, though other sensory interfaces, such as wearable devices, may be worth considering (depending on the context). Choosing an object of study that belongs to one's own culture can be easier to prevent misinterpretation.
2. Read extensively about the historical, cultural, visual, and/or religious context of the object of study to gather background data. This allows a more informed and nuanced understanding of the interaction with the object.
3. Select an approach—comparative, iconographic, symbolic, archival-based etc—or a combination of these. An understanding of which type of contextual data is available and relevant for this digital object can help.
4. Run a detailed visual and/or material analysis of the object of study, from the more obvious features to the more minute details. Note everything there is to be seen and everything that is missing, even if it does not seem relevant at first. Having a clear

research question is important so as not to miss important data. Identifying the point of view of and context of the user(s) interacting with the object of study is key to guide this examination.

5. Sort through the data gathered from the analysis and link it back to the background data and literature to make meaning out of it. This requires sufficient data in order to form an overall argument. Prioritising visual elements that are the most noticeable or important to the user(s) identified is key to create a coherent analysis.
6. Analyse the context of display, use, circulation, and any possible interaction the object of study elicits, according to what brings further light to the analysis. In this chapter, not only do I analyse the French online admissions user interfaces, but I also consider the visual representation of Parcoursup in the national press, as the general reception it depicts can inform us on how university candidates might read the Parcoursup interface.
7. Cross-reference each main point made from the visual analysis with other sources, depending on the background data available, e.g. other user interfaces, official reports and speeches, current legislation, news articles. This ensures the object-based analysis is rigorous and convincing.
8. Refine each point made through the analysis and repeat the process as many times as necessary until you have enough material and a coherent argument to answer the initial research question.

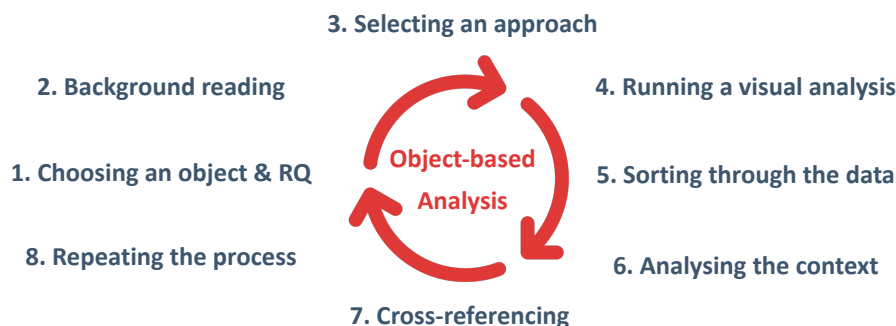


Figure 6.2: 8-Step Object-based Analysis Methods

6.3.2 Object-based analysis helps to better account for the visual rhetoric around transparency

6.3.2.1 The Capability approach applied to algorithmic system transparency.

Object-based analysis has the potential to complement HCI and other design disciplines in unveiling important transparency mechanisms in the interface design of algorithmic

systems. To illustrate this, I apply a capability approach to the transparency regimes and visual rhetoric of Parcoursup. Sen’s concept of capability can be summarised as the realistic means for individuals to choose and access goods they value, such as food, healthcare, or education [308, 307]. This framework has been highly influential to measure level of human development of countries [222, 90]. Since Parcoursup centralises access to higher education in France, the capability approach is relevant here; in this context I consider analysing the system’s interface design and its impact on youth’s understanding of the university admissions system. I here define *capability* as the potential combination of states and activities that enable a teenager to reach the feasible functioning or goal of being accepted into a French university. In line with Gandjour and Nussbaum’s works [129, 254], I also introduce the concept of *sub-capabilities*, i.e. capabilities related to the individual, the system, or their interaction that enable the candidate to reach the main capabilities described below.

Fig. 6.3 summarises some of the main capabilities university candidates need in order to successfully reach the functioning: being accepted into university. These include applying to university, understanding the university admission process, understanding each university’s entry requirements, and meeting the admission process and university requirements. These are all currently centralised online through Parcoursup in France. Based on recent literature in education and technology [243, 388, 86, 6], I identify that teenagers mainly need the following sub-capabilities:

- i. the candidate has a device to access the system, a stable Internet connection, and is e-literate
- ii. Parcoursup is usable on mobile devices: 94% youth in France owns a smartphone in 2021 and over 78% use it to go online. And yet some struggle to use online administrative services, so making Parcoursup usable for teenagers on smartphones is key to facilitate their applications [173, 56, 289]
- iii. the candidate understands how to use the Parcoursup app or website
- iv. they understand how using Parcoursup influences the admission process and outcomes

The issue of usability for students and their understanding of a specific online service has been raised by Zou et al. and Ahmad et al. [388, 6], particularly during COVID-19 lockdowns, when students’ education heavily relied on daily use of such services. Students’ access to online education services has more generally been quantified by The United Nations [243], and Daniel [86] summarises some key challenges and solutions going forward. Whilst (i) refers to the candidate’s socio-technical context, (ii-iv) relate to interface design elements and thus can impact the effectiveness of transparency mechanisms in Parcoursup.



Figure 6.3: University applicants' key capabilities to enter university in France today

6.3.2.2 Object-based analysis helps to assess how interface design influences the capabilities related to transparency.

HCI, UX, UI and interaction design enable us to assess some of the above-mentioned sub-capabilities successfully: for example, the sub-capability (ii) Parcoursup being usable on smartphones is a matter of UI and UX design and §6.5.2 proves how Parcoursup fails at this. However, I argue object-based analysis can contribute to better accounting for the role of the Parcoursup interface in mediating teenagers' capabilities to enter higher education. It contributes to the critical investigation of the sub-capabilities directly related to algorithmic system transparency for university candidates, i.e. the candidate's understanding of (iii) how to use the Parcoursup app or website and of (iv) how using Parcoursup influences the admission process and outcomes:

- The sub-capability (iii) understanding how to use the Parcoursup app or website comprises, I argue, two main sub-capabilities: Parcoursup being usable, and its interface design aligning with users' cultural background and online habits – the latter also influences the sub-capability (iv). Interaction and UX design already offer effective methods to assess Parcoursup's usability [47]. However art history is better suited to critique the alignment of Parcoursup interface design with users' cultural background by casting light on the cultural elements that influence the system's evolution. Section §6.5.1 demonstrates how both APB and Parcoursup do not tap into youth's cultural references, and by failing to do so, the information they provide on how to use the system may not come across as clearly to teenagers.
- Similarly, the sub-capability (iv) how using Parcoursup influences the admission process and outcomes can be divided into: using explainable models within the admission process (i.e. not black-boxes), providing contextually appropriate information [69, 250] about the system to the candidates, and providing such information in a meaningful format to them. Although text-based research methods can be used to

assess whether the model employed is explainable – note that there is currently not enough information publicly available to assess whether the Parcoursup algorithmic model is [127], an inter-disciplinary approach focussing on visual rhetoric is needed to assess whether the formatting of such information is contextually appropriate and understandable for teenagers in France. Section §6.5 illustrates how object-based analysis can contribute to this critical inspection by replacing the Parcoursup interface design within the context of its public reception. Indeed, Fig. 6.13 confirms the system seems opaque to candidates.

To summarise, art history, and object-based analysis in particular, can assist in better accounting for the visual rhetoric around transparency with regards to university candidates’ understanding of how to use Parcoursup and how using it influences the admissions process and their university admissions outcomes. In this sense, it offers an opportunity for inter-disciplinary, critical research on algorithmic system transparency and its visual rhetoric on the user interface that goes beyond well-established disciplines such as UX design. I now introduce the historical background of APB and Parcoursup before visually comparing three key visual representations that reflect possible issues of transparency in their respective designs, and the evolution between the two interfaces (§6.5), in order to show object-based analysis methods applied in practice.

6.4 APB and Parcoursup: historical background

Like its predecessor Admission Post-Bac (APB), the current French university admissions system Parcoursup has been greatly criticised. In line with art historical methodologies, I first provide some background to the development of French admissions systems, to contextualise Parcoursup and the calls to make it more transparent.

Access to higher education in France has been a right by law since 1984 [207]. However, no design of the process to match students and university courses has yet achieved consensus, and problems with the system have a dramatic social impact: over 840,000 university candidates used Parcoursup in 2021 [200]. Parcoursup was designed and deployed in 2018 by the French Ministry of National Education as a centralised online system granting access to higher education on a national scale. The general public had rejected APB through significant strikes, media and official backlash. Moreover, French laws had been passed to increase the regulation of new technologies in 2016 [208].

Since its launch in 2018, Parcoursup has been at the centre of public debates. Like APB [278], it has been negatively assessed by both press and academics, criticised for the unfairness and lack of transparency of its algorithms and overall procedure [238, 41], but also for its detrimental impact on candidates’ wellbeing [288]. Despite the creation of an Ethics and Scientific Committee for Parcoursup (CESP) prior to its deployment [210],

which was (amongst other tasks) charged with improving transparency, issues regarding its opacity have continued to be raised by various stakeholders [92, 237].

Until now, critical assessment of Parcoursup has been conducted mainly by social scientists, where I see publications on the system’s role in equal opportunities, individual’s academic career choices, and social diversity in French society [126, 49, 335]. Employing a socio-technical approach with art historical methods, I consider the design shifts associated with the introduction of Parcoursup together with legal sources, official reports, and media coverage to assess the implementation of transparency on its user interface.

6.4.1 The Introduction of algorithms in French university admissions

Before analysing APB and Parcoursup in detail, the question of whether an online system was needed at all should be considered [67]. Until 1990, French universities admitted students on a first-come, first-served, basis at in-person admissions days. As archival videos show, barely any technology was used [4], where candidates and parents would queue in the street for hours, sometimes camping in tents. Each university ran its own admissions, and French law forbade any kind of selection when allocating candidates to higher education [207]. The system was tedious and inefficient – in 1987, a youth testified at the entrance doors of the Sorbonne that since there was already a queue of 400 people for degrees in law or languages, he had decided to come and apply for medicine instead.

In 1987, a survey of high school leavers’ wishes was deployed via Minitel (pre-Web home terminals used to access various online services through telephone lines in France since 1980). With this RAVEL system, Parisian students could specify three choices of university course, providing the first centralised statistics on student preferences [4]. However, the RAVEL system had drawbacks in fairness and transparency. Firstly, it was introduced as a “census,” leading some teachers and school leavers to feel deceived when they learned that rather than a survey, RAVEL was effectively an admissions system. Secondly, students were selected based on their home address being close to the university, since academic selection was illegal, leading to what some described as geographical discrimination. This led to a sense of unfairness, further fuelled by the fact that school leavers outside the Paris area were excluded from the RAVEL system.

While RAVEL was being turned into an online version by the early 2000s [4], APB was developed in parallel. First tested in 2002 for allocating students to prestigious engineering courses, it was launched nationally in 2009. This was the first system to introduce algorithms into the admissions process, including a ‘stable marriage’ algorithm, i.e. an algorithm that gives a stable match between students and university courses given their respective ordering of preferences, and a ranking classifier to rank applicants whilst

preventing gaming [128, 80]. In terms of visual presentation, APB is mostly text-based, with its home page featuring four large icons reading (Fig. 6.1): ‘What is it?’, ‘For whom?’, ‘When?’ and ‘How?’[143]. I discuss the APB interface in detail below in comparison with Parcoursup.

This brief history shows that transparency issues were already central to the design problem and public debate regarding university admissions, with a series of technological solutions deployed to address them. Seen in this context, the shift between APB and Parcoursup is not new, though it was the first time introducing greater levels of transparency over the online admissions system was seen as a means to redress the information imbalance candidates (and the wider society) had been facing. Yet both interfaces inherited the challenges from this socio-technical heritage. Whilst there was a general consensus for change to the status quo, the design strategy chosen to replace APB was not obvious.

6.4.2 Political drivers: design for transparency?

Parcoursup was introduced following growing public distrust of APB. With more candidates applying to university every year (700,000 school leavers in 2016) and APB selection not based on academic criteria [278], popular courses had been oversubscribed for years [206]. After 40,000 additional candidates enrolled in APB in 2017, a national scandal resulted from random selection at large scale [205]. This was partly due to general university funding problems, but also a design problem since universities did not have the legal and technical means to apply selective criteria. In a blog article titled ‘The APB Scandal’ published on 12 July 2016, the popular economist Thomas Piketty called on the government to reform APB:

the attitude of the Ministry of National Education on this matter is simply outrageous. It maintains an extreme opacity about the criteria used in the software and only shares information bit by bit, whilst regularly promising more transparency, never implemented in practice. (...) This lack of transparency can no longer last.[278]

He compared the French system to other countries, saying:

Like in Germany and Sweden, the principle of free higher education is well established in France. But this does not exempt us from implementing a fair and efficient system for course allocation. Let us not waste the chances this alternative model grants us by replacing private arbitrariness with public opacity. [278]

There was also an increasing gap between French law and the APB system. Transparency is advocated by the OECD as a central value of democracy and trust [150], and

French government policy had committed to increased transparency [256]. The French Public Court of Audit reported on 10 October 2017 that APB was not compliant in terms of transparency and had to be reformed [80]. Furthermore, although French law had banned selection to access higher education since 1984 [207], debates about how to implement this in practice had previously arisen, for example with the abandoned Devaquet draft act in 1986 that was meant to increase universities' autonomy, including the possibility to introduce tuition fees and their own selection processes [207, 362]. Alternative selection processes based on the candidate's postal code (RAVEL), their academic district, or at random (APB) have gradually been introduced, none of which were deemed fair nor transparent. Despite legislation addressing socio-economic inequalities, like a 2013 law reserving some offers in the most prestigious courses for candidates from underprivileged schools [278, 40], or general regulation of technology and personal data such as the 2016 Digital Republic law, such legislation appeared to have little impact on APB [209]. Furthermore, APB did not abide by European GDPR (General Data Protection Regulation) standards for transparency, which came into force in 2016. For instance, APB code had not been made public despite several calls for it, and in 2017 the CNIL (French Data Protection Regulator) took action against the Ministry of Education [77].

As a response to public and legal concerns about transparency and accountability regarding university admissions, the French President Emmanuel Macron launched Parcoursup shortly after being elected. On 28 September 2017, the Minister of Higher Education announced the new Ethics and Scientific Committee for Parcoursup (CESP):

The platform that will succeed to APB will be backed by an ethical committee, which will include high level academics and will check that this framework is strictly respected. It will also oversee that the platform working rules are perfectly transparent. [88]

The committee has been required to publish its findings in annual audits. These precautions demonstrate the government's awareness of concerns over transparency in algorithmic decision-making. A press release from 7 February 2018 states the political drivers for CESP, and the mission to bring more transparency through the launch of Parcoursup:

This commitment marks the government's will of carrying out transparent policies. Setting up this committee therefore constitutes one of the answers to the criticisms about APB for being an opaque system, 'without any human intervention,' and in which the decisions were dictated by non-transparent rules, ruling over the algorithm. [88]

The above sets out the requisite context as is necessary for undertaking an object-based analysis. The following sections investigate how this policy of transparency plays out on the

interface level, through visual comparison of APB and Parcoursup based on art-historical methodologies (§6.2).

6.5 Visual analysis: a tool to assess System transparency in user interfaces

Interfaces are essential objects of study for assessing an algorithmic system and how system transparency is designed for users. In the context of APB and Parcoursup, visual elements can influence university candidates capabilities to enter higher education by impacting their understanding of how to use the user interfaces, and how doing so affects the admissions process and outcomes.

I acknowledge that if no information about the systems and admission processes are given to the candidates, then the impact of the Parcoursup and APB interface designs is limited in terms of transparency for users. I therefore took the following approach: if some type of information over the systems is provided to them, then *the way in which this information is visually communicated to the users through the interface matters*, as it can affect the teenagers' understanding of such information and, consequently, their capability to use the systems to enter higher education. This section gives practical examples of how object-based analysis can be applied to assess whether transparency is effective on the interface level. Note that some visual rhetoric of Parcoursup evolve whether you access the website before or after the admissions deadline has passed (§6.5.2) and that this analysis has been conducted in August 2022.

First, I compare the Parcoursup and APB logos, as visual symbols of the whole systems: as icons that aim to encapsulate the essence of the systems, admissions processes, and user interfaces, they synthesise their respective approaches towards university admissions and set the tone for the candidates' reading of the interface designs. I then scrutinise the login pages, which are important as they constitute the entry point to the systems and, if poorly designed for teenagers, can hinder access to further transparency information; and finally the admissions page, a key feature in the process, where candidates directly interacts with the algorithmic decision-making aspects of the systems. The goals of this section are two-fold: (i) to show the application of object-based analysis in practice, and (ii) to evaluate whether the goals of a more transparent system have been realised or not considering its visual rhetoric and public reception.

6.5.1 Logos: failing to align with users' visual culture can hinder their understanding of the system

Whether a user interface can successfully communicate meaningful information over its system with its audience depends on how contextually appropriate this information is and how closely the interface design taps into the user's cultural background. Many dimensions come into play in the context of making Parcoursup more meaningfully transparent for candidates, e.g. choosing an algorithmic model that is explainable. I argue that in order for teenagers to understand how to use Parcoursup and how using it influences their university admissions, the interface designers need to speak to them in their own visual and cultural language – assuming some type of information to do so is provided in the interface. Since they visually represent the systems as a whole, I start by considering the logos of APB and Parcoursup to illustrate the impact of the interfaces' visual rhetoric on transparency for users.

A close comparison of these logos (Fig. 6.4, 6.5) reveals how political drivers are visually performed through the design process, in the same way that performative speech effects social action [20]. Neither APB nor Parcoursup logos make specific reference to the visual culture and language of Generation Z, thus failing to include its primary intended users into the systems (§6.5.4). This can lead teenagers to feel alienated from the whole procedure, as confirmed by Parcoursup's public reception (§6.5.4). Consequently, it can negatively impact candidates' capabilities to understand how to use the systems and their impact on the admissions outcomes, as well as on the systems' transparency goals.



Figure 6.4: Admission Post Bac Logo [Trademark], via Wikimedia Commons [link].

The APB logo (Fig. 6.4) makes no visual link to either education or the youth culture of the candidate users. The thin capital letters 'Admission,' in monochrome light petrol blue on a white background, are followed by 'Post Bac' in smaller bold, white capitals set into a rectangular bubble with rounded edges. Compared to popular culture in France, that bubble could be read as an attempt by the government to cite francophone comic books, although the famous *bandes dessinées* of France are associated with the childhood of baby boomers rather than Gen Z, who are more likely to identify with the speech bubble style of Japanese manga [119, 21, 114]. The flat design of the APB logo mostly resembles digital services logos of its time [368]. The uppercase letters denote an official message, the blue evokes online service platforms of the 2000s, and the text box recalls administrative forms of data exchange. This text-based logo and its text box can recall the paper forms of earlier university applications (§6.4.1).

Whilst the Parcoursup logo (Fig. 6.5) clearly differs from its APB predecessor, it

similarly fails to reference youth-oriented visual discourse. It does, however, reflect the newly elected government’s politics on regulating access to higher education [97]. The single lowercase word merges two academic terms: ‘parcours’ and ‘(études) sup(érieures)’, corresponding to the English terms ‘academic path’ and ‘higher education’.



Figure 6.5: Parcoursup Logo [Trademark][link].

This lexical juxtaposition is visualised by the combination of dark bluish grey for ‘parcours’ and bright red ‘up’ with the ‘s’ being vertically split into both colours, as if to symbolise a pathway with this curved line. Three stylised arrow heads (two red ones framing a grey one) head toward the upper right, with the last arrow meeting the top of the first letter. They can be read as symbolising moving up in the world through education, and/or a mouse reaching ‘parcoursup’ and nudging the user to click on it [328]. The integrated motto ‘Enter higher education!’ reinforces this latter reading, where clicking to enter the Parcoursup portal becomes a metonym for access to higher education overall – something the capability approach also reveals (§6.3). This metonymic action emphasises the dominance of the French state, since Parcoursup is virtually the only way for high school leavers to access university today [233].

However, even stronger than references to education is the visual connection between the Parcoursup logo and major supermarket brands. Although the colour reflects the French flag often set near it on Parcoursup pages (e.g. Fig. 6.1), it also resembles earlier logos of famous French supermarkets: the 2013 Monoprix logo for its red and detached font, the pre-2009 Carrefour logo for its two-colour code and round font, and the Franprix logo of 2006-2015 for its layout with the visual preceding the word and red lowercase text (Fig. 6.6).



Figure 6.6: Monoprix (link), Carrefour (link), and Franprix (link) Logos [Trademarks].

These logos were already out of use when the design of Parcoursup started in 2018, but had been ubiquitous in French urban spaces in the early 2010s. In responding to distrust of APB, the goal was to rebrand algorithmic admissions. My visual comparisons show how the Parcoursup logo appears to address the middle class, middle-aged, and urban adult audience who regularly shop in French-branded supermarkets. It does not seem designed to appeal to French teenagers who do not see such supermarkets as either fashionable

or aspirational. The choice of logo suggests to this audience that university candidates themselves might not be considered by the French Ministry of Education as the agents or actors of their own academic careers [201]. This may be partly because the Ministry has a monopoly that does not need to attract users in the same way as a private brand. Perhaps by appropriating familiar supermarket logos, the Parcoursup logo was designed to give an air of solid respectability, just as the possibly aspirational message of the rising arrows is instead dominated by the utilitarian suggestion of where to click.

In the context of government monopoly, brand recognition through logo design does not primarily serve to attract customers, but rather to reinforce public trust in the French higher education institution. The first Macron government had made a commitment to move away from approaches to algorithmic decision making that had led to public distrust in APB, and had a political perspective from which the public were expected to have greater trust in private commercial companies [329]. These drivers might explain why this logo does not resemble conservative single-color logos of public sector bodies such as the social security agency and its online portal (Fig. 6.7). However, since using Parcoursup is



Figure 6.7: Assurance Maladie and Ameli.fr Logos [Trademark] [link].

virtually required to access much higher education and the French state has the monopoly of allocating students to public universities (§6.4) [207], choosing a logo that does not align with today teenagers' visual references is not benign: it seems to indicate that Parcoursup is not tailored to university candidates, despite them being its primary audience, which points towards a discrepancy between the calls for greater transparency for users and the system's branding.

6.5.2 Login pages: visualising constraint

The object-based analysis of the APB and Parcoursup logos seem to indicate there is a gap between the calls for increased transparency over the admissions systems and the Parcoursup interface design. I now show how the visual rhetoric framing the information given to teenagers in the login pages undermines its clarity and how that can affect their capabilities related to transparency. Similarly, the purpose of the APB login page (Fig. 6.8) is to open up access to higher education via the Internet. However, my visual analysis reveals its login page creates constraints that can erode users' understanding of how to use the university admissions website.

The visual design of the APB login page emphasises rectilinear lines, with square boxes and delineated rectangular areas. The only round form is the standard pictogram of an official hotline. Inconsistent fonts, colours and pictograms communicate awkwardness rather than providing information clearly. The welcome message ‘Bienvenue sur Admission PostBac’ in bold, black uppercase hovers in empty space rather than being spatially connected to the rest of the page.

Figure 6.8: APB home page [Public Domain], via Wikimedia Commons [link].

When applying this critical reading, one might argue a login page is purely functional, which is why no thought was given to making it visually welcoming. However, this lack of thought in itself sends a message – one that contradicts the supposedly welcoming text. An inconsistent formatting of information that furthermore does not align with the users’ visual language (§6.5.1) can hinder the clarity of the interface design. If not granting easy access to the admissions system, the login page can reduce users’ access to further information about the system and its impact on the university admissions outcomes, i.e. the capabilities directly related to transparency (Fig. 6.3). Although public services may not need to invest in design to compete with direct competitors, they still have obligations of usability, inclusion, and accessibility, requiring some professional design attention [58], and should convey an air reliability.

The Parcoursup login page (Fig. 6.9), while more visually consistent, still fails to welcome students – in fact, it visually hinders access. Instead of a welcome message, the functional elements are squeezed at the top of the page between the bureaucratic

Figure 6.9: Parcoursup Login Page in August 2021 [Trademark], via Parcoursup website [link].

stages of the admissions process, which is dominated by a bright red warning banner (an unusual colour choice for the ‘help’ link replacing the APB hotline), and at the bottom a legal footer with more than 10 links to other sites. The result is visual constriction and stress relating to the admissions requirements [288]. Note that the visual rhetoric on the login page evolves according to whether you access it before or after the final admissions deadline has passed, e.g. the banner turns yellow after the deadline. While the design is conventional, resembling templates for login authorisation on packaged design sites such as Wix.com [367], reliance on standard visual trends raises the question of whether designers recognise the function of this service in granting transparent access to a public system and a public service. The bold colour-block horizontal strips defeat the purpose of the login screen by drawing the candidate’s eye away from the barely visible login box. This contradiction between access and design is even more striking on a smartphone screen, where the lack of responsive layout makes the login information and entry boxes become invisible and impossible to use without zooming in (Fig. 6.10). The fact that the phone layout has not been considered gives a particularly strong message to Gen Z users given that 94% own a smartphone in France today and many use it to go online: their online usage and preferences do not seem to be prioritised.

In all, my analysis of the logos uncovers a gap between the calls for greater transparency over the admissions system and the Parcoursup interface design. This discrepancy is also visible when scrutinising other key stages of the admissions process, starting with the login pages. A logo that fails to align with its audience’s visual culture can impact their overall impression of the interface; likewise, as the first point of access to any further

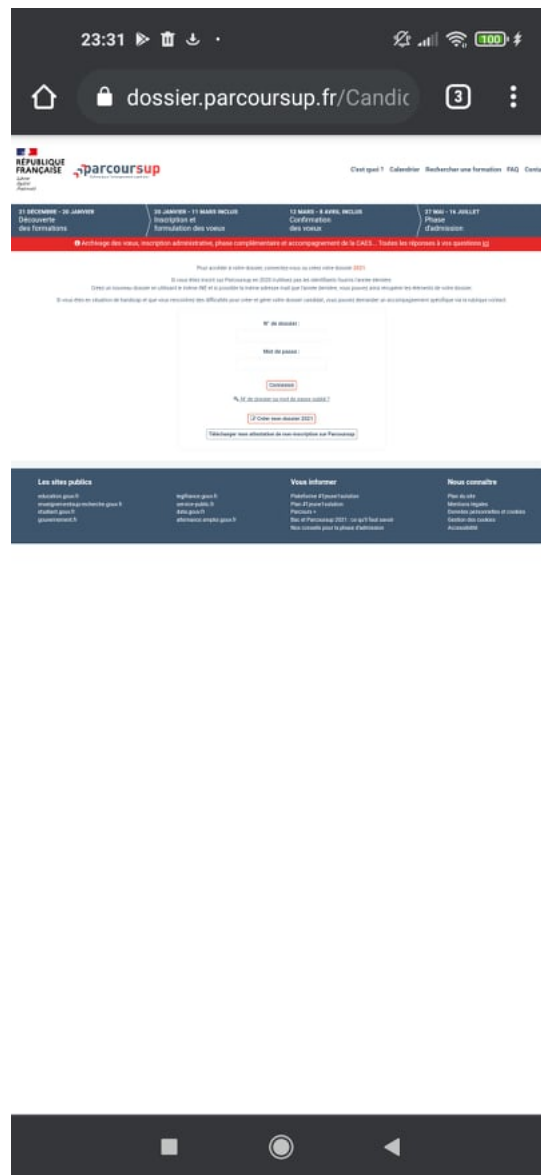


Figure 6.10: The Parcoursup login page on a smartphone in August 2021 [Trademark] [link].

information about the system, a lack of visual clarity in the login pages matters: together they can affect the candidates' capabilities not only to understand how to use the system, but also how using it can influence their admissions outcomes. Indeed, the interface design of login pages can influence candidates' reading of the following Parcoursup pages and the transparency information they might provide. It is therefore essential for the Parcoursup login pages, among others, to better align with teenagers' cultural language and online habits to increase transparency for them over the admissions system.

6.5.3 Admissions pages: allocation transparency

I have demonstrated how the logos (by which candidates are initially directed to the system) and login pages (their first substantive interaction) include design elements that can influence to the effectiveness of algorithmic system transparency mechanisms. Although Parcoursup was designed specifically as a response to public distrust in APB and public demands for greater transparency, I have shown that the design of these initial elements does not appear to work towards such goals, as they do not align with the cultural and online background of its core audience of Gen Z.

The final element of my analysis considers the visual presentation of the algorithmic decision itself, in the admissions pages of the two systems. This page aggregates the communication of the admissions outcomes, and allows the student to respond to the allocations. The visual design of the admissions pages represents the climax of stress and anticipation for candidates, but also crystallises the issues of transparency and accountability associated with the two systems. Through the object-based analysis, I highlight what is *not seen* on the interface, i.e. the lack of information candidates receive about the algorithmic decision-making process, but also how they perceive the whole systems as opaque (§6.5.4).

The algorithmic basis of the two systems differs. As already described, APB applies a ‘stable marriage’ algorithm and a ranking classifier to rank applicants whilst preventing gaming [128, 80]. Parcoursup uses an adapted and slower version of the stable marriage algorithm (that runs without students initially providing an order of preference); a machine-learning-based decision-support assistant for universities to select students (note there is no official information publicly available on these local algorithmic models); and an adjustment algorithm to boost the ranking of applicants receiving social benefits, in order to meet quotas for under-represented groups [219].

A screenshot of the APB admissions page (Fig. 6.11) shows two tables: the upper offering buttons for candidates to give their replies, and the lower showing the list of courses they applied to, with the corresponding offer from each university (‘Situation’). In this example, the candidate has been offered a place in their first choice application (‘Formation proposée’), meaning that the remaining applications have been withdrawn automatically. Other possible offer values might be ‘waiting list’ or ‘rejected,’ in which case the otherwise empty fourth column would include a button to withdraw individual applications. No further opportunity is provided for candidates to request further explanation – the only interaction option in this example is to click one of the two buttons ‘definitive yes’ or ‘I withdraw all my applications’, providing the user with very little flexibility and autonomy [59].

Despite the change in design objectives between the two systems, the interaction afforded by the Parcoursup allocations page (Fig. 6.12) remains surprisingly close to that

INSCRIPTION

SCOLARITÉ

VOEUX

ADMISSION

PROPOSITIONCOORDONNÉESJE RENONCE À TOUS MES VOEUX

PROPOSITION D'ADMISSION

Votre admission est proposée à la formation suivante :

Université Claude Bernard Lyon 1 (Villeurbanne) - Médecine, Pharmacie, Odontologie (dentiste), Maïeutique (sage-femmes) - PACES (Première année commune de Santé : Médecine, Pharmacie, Odontologie, Maïeutique, Kinésithérapie, Ergothérapie) - vœu n° 1 .

Vous n'avez pas encore répondu à cette proposition. En l'absence de réponse, vous serez démissionné(e).

Vous souhaitez accepter définitivement la formation proposée, cliquez sur le bouton ci-contre

Oui Définitif

Vous souhaitez renoncer à toutes les formations (y compris celle qui vous est éventuellement proposée) cliquez ci-contre

Je renonce à tous mes vœux

- VOTRE LISTE DE VOEUX ORDONNÉE -

Vœu	Formation	Situation	
1	Université Claude Bernard Lyon 1 (Villeurbanne) - Médecine, Pharmacie, Odontologie (dentiste), Maïeutique (sage-femmes) - PACES (Première année commune de Santé : Médecine, Pharmacie, Odontologie, Maïeutique, Kinésithérapie, Ergothérapie)	Formation proposée.	
2	Université Claude Bernard Lyon 1 (Villeurbanne) - Portail portail Sciences de la Vie et de la Terre	Démission car vœu 1 proposé.	
3	Université Lumière-Lyon 2 (Lyon) - Psychologie Psychologie et Complément de majeure Psychologie	Démission car vœu 1 proposé.	
4	Université Lumière-Lyon 2 (Lyon) - Psychologie Psychologie et Sciences de l'Education	Démission car vœu 1 proposé.	
5	Université Claude Bernard Lyon 1 (Villeurbanne) -	Dossier non parvenu.	

Figure 6.11: APB Admissions Page. Screenshot posted on a candidate’s blog in 2015 (link). The page includes 2 tables with the first offering buttons for candidates to express their final replies and the second below summarising the ordered list of courses they applied to.

of APB. The image of a university candidate viewing the system admissions page in 2018 is widely circulated, highlighting how the limited user interaction with this page has become a metonym for the Parcoursup system as a whole. In the Parcoursup allocations page, the two APB tables have been merged into a single table showing application outcomes in a non-ordered list, with a new right-hand column headed ‘yes if’, where two buttons allow candidates to make a positive or negative decision for each application. However, with regard to transparency, there is still no interaction option by which candidates could understand or investigate the selection process and its outcomes, which further illustrates the importance of the detailed and critical investigation afforded by the object-based analysis I have experimented with here.

The comparison of admissions pages further indicates that transparency aims have not been met on Parcoursup, given that no major improvement can be seen over the APB design. My visual analysis leads me to ask how the universities and public services this system represents can be made accountable to civil society. Moreover, how can such designs build trust in other new technologies for automated decision-making, or promote, assess, and regulate transparent and sustainable design standards for online services and algorithmic systems in general?

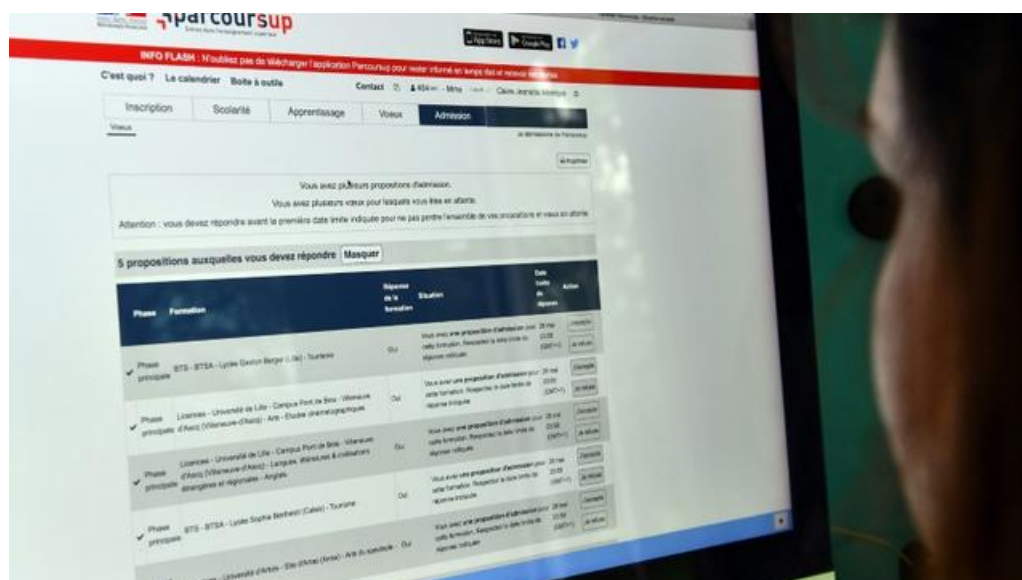


Figure 6.12: A Candidate viewing their admissions in Parcoursup on a desktop in May 2018. Photograph by Denis Charlet/AFP (link).

6.5.4 Preliminary conclusions and Parcoursup's public reception

Based on this analysis, it is clear that the demands for transparency and for a change of system were justified, given the visual opacity of APB. However, my analysis also shows that Parcoursup, when considered holistically in terms of visual rhetoric also failed to deliver effective transparency, despite this being a key reason it was commissioned. The analysis of logos shows both interfaces fail to directly target its key audience, generation Z. The spatial arrangement of APB and Parcoursup login pages equally fails to welcome students, with Parcoursup giving emphasis to procedural aspects, without consideration to access from smartphones. Finally, little information was provided regarding the allocation process and algorithmic system on the APB admissions page, and the Parcoursup design did not change this status quo.

The media response to Parcoursup confirms my own conclusions from object-based analysis of its visual rhetoric: the public generally considers Parcoursup as failing its transparency goals. As seen from satirical images, for instance, a 2021 cartoon by Aurel (Fig. 6.13) points at this gap between official discourse about transparency and what this actually entails for candidate users. The cartoon depicts the encounter between a backpack-wearing student and a civil servant behind a Ministry counter. The latter says through a glass pane: 'A trap?', 'Which trap?', 'There is no trap'. Rather than being transparent and welcoming, the wall behind them labelled 'PARCOURSUP' is covered in spikes and barbed wire. By replacing Parcoursup in the context of its public visual reception, my method further highlights how it fails to communicate information in a meaningful and contextually appropriate visual format to university candidates, so as to

empower them to fully understand how to use the system to enter higher education.

This case study has therefore demonstrated how object-based analysis is particularly relevant to assess the effectiveness of system transparency: not only can it point at the potential gaps between transparency policy and its design on the interface level, but it can also highlight the interplay between visual rhetoric of transparency and its public reception, therefore encouraging the design of more transparent and accountable systems.



Figure 6.13: Parcoursup, a Trap?. Cartoon by Aurel. Published in Le Monde website on 2 July 2021 ([link](#)). The man at the Ministry desk says in French: ‘A trap?’, ‘What trap?’, ‘There is no trap’.

6.6 Discussion

Through this case study, I have demonstrated the closeness and importance of the relationship between critical analysis for transparency assessment (regarding the broad concerns in public-facing algorithms systems) on the one hand, and critical analysis of the ways that transparency might be visually effected on the other. Indeed, through an interdisciplinary socio-technical approach, I have shown that the interface is an essential object of study for assessing an algorithmic system, since that is the point in which people use and engage with it, and APB and Parcoursup offer a highly suitable design case study for an investigation of how these aspects come together. Due to the growing calls for increased transparency surrounding the French admissions systems, I analysed the interfaces of systems to assess their transparency-related affordances. I show how the object-based analysis I have experimented with here is insightful to critique the effectiveness of transparency mechanisms (§6.5.4). However, I acknowledge that interface design

is only one lever to increase transparency for users, e.g. the CESP annual reports [391] for example also work towards this aim (§6.4). More broadly, I argue for a distinctively holistic socio-technical approach with regard to how transparency and visualisation are understood, assessed and critiqued in design undertakings.

Towards this, I have presented a socio-technical approach that focuses on the visual rhetoric involved in the transparency dimensions of user interface design but is distinct from, yet complementary to conventional methods of analysing transparency. I argue that this method is also helpful in contexts where user studies might not be appropriate or practically doable. For instance in our case study, it has been widely reported that many university candidates find their experience interacting with Parcoursup very stressful, sometimes even leading to anxiety, so I do not consider ethical to ask them directly about it through a user study. Instead, I have shown that a comparative object-based analysis is sufficient to outline clear gaps in the APB and Parcoursup interface design in terms of algorithmic system transparency. Moreover, given its critical attention to the meaning and context of the object of study, the added-value of applying art history to critical assessment of interface design is not to be confused with that of aesthetics nor UX design. Whilst aesthetics is closely connected to philosophy, focussing for example on discussions about taste, art history on the other hand is primarily concerned with providing historical, visual, material, symbolic, and/or socio-cultural interpretations of artworks. In other words, the general purpose of art history is to *situate and identify the possible meanings of objects in context and how its beholders have interacted with it*. It is **not** interested in whether an object is “beautiful” nor “appealing” to users. Similarly, art history is not concerned with UX design goals, such as usability. Although extending the methods and tools to make technology more useful, usable, and desirable is essential, and attracts a lively research community [1], I argue that art history can serve other aims: it can help critique user interface design by better contextualising and accounting for the impact of visual culture on its design, and in particular assess its transparency dimensions. As such, I propose and argue the urgency for such methods to be taken up and used as issues and concerns regarding transparency in algorithmic systems are growing.

Moreover, my method can be successfully applied to the study of virtually *any* interface design to assess the implementation of its transparency regimes. Many current art historians are investigating non-art objects, e.g. objects from our daily lives. This coincides with a more holistic vision of cultural heritage that has led to the creation of over 300 ecomuseums since 1971. Unlike traditional art museums, ecomuseums do not focus on specific objects but rather aim to highlight a local culture and community through its objects [279]. To better account for this wider definition of art and heritage, current art history is thus increasingly entangled with anthropological theories [287, 293]. I follow a similar interdisciplinary trend in applying an object-based analysis to non-art user interfaces, such

as APB and Parcoursup.

Where systems such as Parcoursup have been designed specifically to address earlier failings of public transparency (rather unsuccessfully, as I have shown), HCI undertakings must also pay critical attention to analysing public reception. Indeed, the lack of visual transparency in the original deployment of the Parcoursup system left others to fill the gap: the most original, accessible, and thorough visual explanations of the Parcoursup algorithms and functionalities have been provided by members of civil society, namely famous French YouTubers such as *Hugo Décrypte* by Hugo Travers [337], *Science Etonnante* by David Louapre [218], or *Bright Future* [127]. Louapre’s video reached over 540,000 views, reflecting the extent of public need for and interest in such. The fact these videos are hosted on a commercial platform and produced by French citizens who have no direct relationship to the design team is telling of the extent to which the Parcoursup interface itself fails to provide transparency through accessible visualisations. The code of the main Parcoursup algorithms – i.e. those that carry out the matching and ranking of candidates with university courses at first, and introduce social quotas in the final admissions stages – has been published on a separate platform, but not the code used by each university for selecting the shortlisted candidates. However, publishing code to a non-technical audience is an example of *illiterate opacity* [69]. Furthermore, as noted in *Bright Future*’s video [127], without an anonymised dataset to test the algorithm, it is impossible to reproduce the selection process and inspect it for potential errors or bias.

Further information has become available through civil society response to Parcoursup, but this cannot replace the government’s own duty for transparency. YouTube videos cannot have the same reach, and may be less accessible to Parcoursup users than if hosted directly on the Parcoursup site. Similar arguments can be made regarding the systems of private organisations, particularly the large scale platforms that increasingly impact many aspects of everyday life.

6.6.1 Future work: integrating object-based analysis into design practice

This chapter seeks to draw attention to the potential for art historical, object-based analysis to assist the design and critique of algorithmic systems. In this way, my work represents a starting point, opening up a number of research opportunities for the community to take forward. In particular, there is an opportunity to consider how the methods can align with and complement current design practices. One way forward is to explore more case studies. Moreover, there is an opportunity to compare and combine the results from object-based analysis with other qualitative or quantitative research methods. Indeed, cross-referencing visual analysis with archival sources is common practice in art history, and so providing

different sources of data to support an object-based analysis of online visual rhetoric can be equally fruitful to justify its value and usability for HCI research. This could result in interdisciplinary approaches that better account for transparency concerns in algorithmic systems.

As far as the case study is concerned, there is scope for further interrogation across education systems and public policy about transparency in a wider context. This could entail, for example, comparing Parcoursup to other European higher education assignment systems, such as UCAS in the UK, in terms of their interfaces, transparency elements, and broader policy context. This comparative method can lead to promote better transparency mechanisms in different contexts - leading to more accountable systems in the education sector and beyond.

Moreover, since art history is the study of how to interpret visual objects, whether physical or digital, it works to highlight the interplay between visual rhetoric of transparency and its public reception. With both the increase of complex algorithmic systems and the increasingly visual nature of digital communications, it is important to consider these aspects jointly when considering the design of more transparent and accountable systems.

6.7 Conclusion and main takeaways

In this final study, I have shown the potential for bringing art historical methods into the transparency assessment of algorithmic systems as socio-technological systems (Assessment Challenge (RQ3), Literature gap 3). In §2.2.3.2, I have explained how focusing on interfaces when doing so can help recentre transparency research on non-expert users and better align system designs with users' information needs. Object-based analysis, as applied to the case study of the APB and Parcoursup systems here, further reveals how critical analysis of interface design can help understand the way algorithmic systems are received by non-expert users. As discussed throughout this thesis, this bottom-up and empirical approach is key to promote meaningful transparent systems for users going forward. Using this critical method, I showed how APB and Parcoursup's branding, access, and allocations outcomes failed to meet the transparency aims that were driven by public discourse – where transparency was supposedly the key driver for the introduction of Parcoursup in the first place. This study thus confirms our description of user interface as an essential object of study for assessing an algorithmic system in §2.2.3.2.

Qualitative methods such as object-based analysis that account for a system's broader social, historical and cultural contexts are relevant for design in general, and can participate to advance HCXAI research in particular (see Chapter 5). Indeed, they are especially relevant in the context of algorithmic system transparency, as they help in determining its appropriateness (see Contextual Integrity in Chapter 3), and how the system may be read

and received, both by non-expert users and general public. Through this case study, with detailed examples of how to apply this method to elements of the user interface, I hope to provide ways forward on encouraging and supporting the realisation of more transparent algorithmic systems. Although I do not claim to have fully tackled the Assessment Challenge (RQ3) in this study, I have opened new research avenues towards this and have started to help fill the Literature Gap 4. I now conclude this thesis by discussing my main contributions and takeaways, as well as the main limitations, and further research avenues to design meaningful algorithmic system transparency for non-expert users.

Chapter 7

Discussion

I now highlight key insights and methods I used in this thesis, before reconceptualising meaningful transparency for non-expert users and discussing the main limitations and further research to improve algorithmic system transparency for users.

7.1 Addressing the four literature gaps

To answer CRQ and improve transparency for non-expert users of algorithmic systems, I have explored the four literature gaps listed in Table 1.1.

7.1.1 Literature Gap 1: Better defining the needs of non-expert users

The first gap I have identified in the human-centred literature on transparency consists in better understanding the information needs of non-expert users. To address it, my studies have investigated the transparency needs of parents in the UK (Chapter 3), and of a specific group of clinicians, i.e. gynaecologists (Chapter 5). In both user studies, I have used a community-based approach to better account for people’s broad environments: first, by recruiting participants among three communities parents in the UK, and then by organising a co-design workshop with a team of gynaecologists in the Netherlands. I have also shown how these types of users have been under-studied and therefore deserve more attention in HCI and HCXAI research. By directly engaging with such users in real-life settings, I have thus stressed that their information needs might vary, even within the same profession, and have called for a more nuanced description of “the who,” [109, 51] when using the label ‘non-expert users’ in HCI and HCXAI research. In doing so, I follow Verma et al’s call to go beyond the ‘one-size-fits-all’ paradigm still present in HCAI research [350]. In better outlining such users’ information needs, I have also moved towards better defining what constitutes meaningful transparency for them, thus providing empirical insights that

can inform future system designs. Indeed, I have explored and reported two different qualitative methods to better account for users' information needs. When comparing the interview study with parents (Chapter 3) and the co-design study with gynaecologists (Chapter 5), I believe the co-design activities proved to be more effective in helping define meaningful transparency for such non-expert users. I would therefore encourage the HCI, XAI, and HCXAI communities to prioritise similar participatory studies, for example by directly engaging with patients as co-designers of CAD tools, as patients represent crucial stakeholders of clinical AI systems.

7.1.2 Literature Gap 2: Exploring the gap between new transparency guidelines and design practices

The second literature gap I have highlighted is the need to explore the gap between new transparency standards and design practices. To help bridge this gap, I have explored the IEEE 7001 transparency standard in relation to designers' practices and understanding of algorithmic system transparency (Chapter 4). Through an online survey and series of interviews with designers, I have shown that the IEEE 7001 standard recommendations might not be effectively implemented in practice by designers, due to a mismatch between their respective approaches. Moreover, this study has highlighted the need and potential to raise awareness about transparency in algorithmic systems among designers: my mixed-method study has contributed to raise awareness about transparency among participants. I thus have demonstrated that the 'principles-to-practice' gap described in the HCI and XAI literature (see §2.2.1) also extends to the latest standards promoting transparency, and my critical insights confirm that designers still need more support to improve algorithmic system transparency for users going forward.

7.1.3 Literature Gap 3: Operationalising HCXAI design recommendations in real-world settings

To further close this 'principles-to-practice' gap, I have tested and expanded the latest design recommendations from the HCXAI literature in Chapter 5. In doing so, I aim to address the third literature gap listed in Table 1.1, i.e. the need to better operationalise such recommendations in real-world settings. I have observed that the HCXAI design recommendations constitute a starting point to discuss meaningful transparency with my selected group of non-expert users, i.e. gynaecologists. However, with a co-design workshop and individual interviews with gynaecologists, I have also shown that HCXAI must better account for clinical teams with different types of gynaecologist users. I have stressed that the timeliness and relevance of the information given to users about an AI-based tool is key for transparency to become meaningful for clinicians throughout the

design lifecycle, and in particular information about the tool’s conceptualisation, role, and purpose provided prior to its implementation in clinical practice. This study thus confirms that more direct engagement with specific user groups can help to better operationalise such recommendations. Towards this, I have provided empirical insights as well as four design recommendations to make algorithmic systems more meaningfully transparent for gynaecologists.

7.1.4 Literature Gap 4: Evaluating transparency in practice

Lastly, I have opened new avenues to research the lack of methods to evaluate (meaningful) transparency going forward (Literature gap 4, see Table 1.1). My study with designers (Chapter 4) suggests that it is still unclear whether the transparency dimensions of an algorithmic system can be measured with a quantified approach, such as IEEE 7001 standard Transparency Levels (TLs). I have thus introduced and experimented with two qualitative approaches towards assessing meaningful transparency in practice. In Chapter 3, I have used contextual integrity as a theoretical framework to benchmark meaningful transparency for digital parenting, based on my interviews with parents. Lastly, I have experimented with a critical and inter-disciplinary method, object-based analysis, to assess a system’s transparency dimension in user interfaces in Chapter 6. I have shown that both these approaches enable to qualitatively describe what constitutes meaningful transparency in algorithmic systems for non-expert users, with and without direct engagement of users. Although they constitute preliminary steps towards tackling the challenge of evaluating meaningful transparency in practice, these two approaches represent my main critical contributions and stress the need for more research in this space. I now discuss the main takeaways beyond the four literature gaps outlined above.

7.2 Summary of main takeaways

I summarise below the following key takeaways from this thesis:

- algorithmic system transparency is important for non-expert users
- there remains many obstacles to improving transparency in practice
- in particular, interfaces need to better account for the users’ broader contexts
- and most user and designer participants do not have a clear understanding of transparency
- however, there are opportunities for closing the gap between transparency theory and design practice by directly engaging with users and other stakeholders

I have shown throughout this thesis that algorithmic system transparency is important for non-expert users and can support a wide range of goals [359, 182]. I have identified that transparency can even support users’ goals that go beyond ethical and responsible AI values such as accountability, or human control (Chapter 2). For example, Chapter 3 shows how meaningful transparency can support digital parenting, while Chapter 6 reveals it can improve access to higher education for university candidates. However, I have also stressed many obstacles to improving transparency in practice, both from designers’ perspective (Tables 4.7, 4.8) and according to users themselves (Chapters 3, 5). In addition to the many challenges already highlighted in the XAI and HCXAI literature, I have for example shown that current user interfaces need to better account for the users’ broader visual and cultural contexts (Chapter 6).

Moreover, I have observed that most non-expert users and designers who joined my studies do not have a clear understanding of what transparency means (Chapters 3, 4), and seldom differentiate between the various types of transparency, which are currently described in the academic literature (Chapter 2). My studies thus illustrate the current gap between theory and design practice, but also the opportunity for closing it by directly engaging with various stakeholders: in particular, users. Such efforts are essential to make sure we better align algorithmic systems provided to users with their information needs, and thus contribute to systems benefiting users and society at large. Towards this, I have outlined both the need and exemplary methods for: i) more collaborative efforts between designers and standards bodies setting transparency requirements and metrics (Chapter 4); ii) better engagement of non-expert users in the design process of meaningful transparent systems (Chapter 5); and iii) better accounting for the communities which influence the users’ transparency needs (Chapter 3) and the broad socio-technical context that influences users’ reception of interfaces (Chapter 6). I now provide a new conceptualisation of meaningful transparency based on my thesis’ findings and contributions.

7.2.1 New conceptualisation of meaningful transparency for non-expert users

As discussed in Chapter 2, previous human-centred research and socio-technical approaches to RAI have stressed that transparency is a *relationship* between designers and users and is *context-dependent*. For example, Felzmann et al. call to “understand transparency relationally, where information provision is conceptualized as communication between technology providers and users, and where assessments of trustworthiness based on contextual factors mediate the value of transparency communications” [120]. In this thesis, I extend this conceptualisation of transparency by showing what meaningful transparency might entail in practice as a transfer of information to non-expert users. First, I have argued that this

requires a more nuanced understanding of non-expert users. Indeed, in Chapters 3 and 5, I show that such conceptualisation needs to account for the various relationships between users themselves (e.g. parents communities, MDTs), the types of communication taking place between them (e.g. the distribution norms that dictate the information flow between the parents I have interviewed in the UK), and the different characteristics that might influence individuals’ needs within a single group of non-experts users (e.g. gynaecologists), for example their mental models (Table 5.5). In doing so, I build upon recent HCXAI research promoting a deeper conceptual understanding of users that moves beyond the “one-size-fits-all” paradigm [350]. I demonstrate this can be achieved by directly engaging users in real-world settings, thus providing methods and examples of how to do so (e.g. co-design activities). Additionally, I use a socio-technical system approach, and claim the visual and cultural background of non-expert users must also be considered when designing and assessing algorithmic system transparency (Chapter 6). In particular, I have shown such considerations are relevant when critically assessing user interface designs to bridge the principles-to-practice gap that has been described in the RAI literature (Chapter 2). As such, I argue they should inform future research in the field.

Moreover, in Chapter 5, I have demonstrated that for transparency to become meaningful, the algorithmic tools themselves require more conceptual understanding. Indeed, this goes beyond any individual XAI intervention or method, and it can, in some contexts, imply separating the concept of transparency from the tool itself. For instance, gynaecologist participants in my co-design activities were unanimous about the need for transparent information to be provided to clinicians *prior to* an AI tool being implemented into clinical practice, in this case for ovarian cancer diagnosis (Chapter 5). In other words, transparency design and considerations in this setting should precede the algorithmic system design, which goes even beyond what the concept of ‘transparency by design’ recommends [121, 227]. Following this logic, this new conceptualisation of meaningful transparency for users would require a reframing of HCXAI by shifting away from explainability towards new, pre-implementation transparency mechanisms. In other words, I argue that the design of transparency dimensions should not be bound and limited to the AI design cycle, and that instead, in some contexts, it might require its own design cycle – see Fig. 5.1 for an illustration of a transparency design timeline in a specific context. While this new conceptualisation requires further research to be strengthened and confirmed, I now comment on some secondary insights from my studies.

7.3 Other contributions to HCI and AI research

In addition to the insights mentioned above, I have also contributed in this thesis to advance HCI and AI research relating to women’s health (Chapter 5) and teenager’s and

families' welfare (Chapter 3, 6) by directly engaging with users in real-world settings. My commitment as a female researcher to serve people by helping to improve how we design systems has motivated this whole thesis, and has been inspired by Peters and Calvo's work on 'design for wellbeing,' i.e. design that supports people's psychological wellbeing, even beyond user experience [277, 58, 59]. In doing so, I build upon their participatory work towards promoting responsible and ethical technology development, as well as the growing body of feminist HCI literature [26, 27, 160], e.g. publications exploring women's bodily experiences and interactions with health and wellbeing technologies [11, 189]. By focusing on transparency for users – in particular to serve accountability goals (Chapter 2)– I thus follow Drage et al.'s recent call to investigate the feminist responsibility of engineers to build responsible AI systems [99]. I now summarise the thesis main limitations.

7.4 Main limitations of the thesis

Overall, the main limitations of this PhD thesis lie in the methodology. I experimented in each of my studies with a different method –i.e. qualitative, quantitative, and critical– which provided me with a rich learning experience. As a result, some studies could have been done differently. For example, the snowballing method to recruit parents within their communities in Chapter 3 proved to be very difficult and significantly capped the sample size. I chose this method to ensure each parent would know at least one other parent from the same community, to better account for the community influence on their digital parenting and transparency needs. Looking back, I could have probably chosen another recruiting method, such as direct mail invitations based on postcodes or children's schools. This would have enabled me to conduct a larger study and likely strengthened my contributions with this study. I therefore acknowledge that other approaches exist, however with this method I made sure the parents we interviewed were all connected to at least another participant, which enabled me to better account for the influence of their respective communities.

Likewise, in Chapter 6, other researchers have encouraged me to confirm my findings with a user study of the APB and Parcoursup interfaces. I did not conduct such a study because I argue the object-based analysis is self-sufficient: in art history, you do not check your analysis of an artwork by asking its author nor the audience whether they agree with it. However, I acknowledge that some might disagree with this stance. I have also demonstrated that there is currently no single method to evaluate a system's transparency level, and thus have suggested different ways that one might approach such empirical evaluation (see Chapters 3 and 6). Other more specific limitations have been discussed in each chapter.

Lastly, I acknowledge that the choice of key challenges and literature gaps I have

explored in this thesis is far from being exhaustive (Table 1.1). For example, in their 2023 systematic literature review, Haque et al. [152] identify 10 research questions and challenges for XAI from a user’s perspective. However, it is not possible to address them all in this thesis. I have therefore prioritised four literature gaps that, I argue, can illustrate the breadth of human-centred research on algorithmic system transparency for users, and provide a starting point to promote more meaningfully transparent systems for users going forward. I now highlight two research avenues on transparency.

7.5 Further research on meaningful transparency

Lastly, I highlight two key research avenues for designing more meaningfully transparent algorithmic systems for non-expert users. I encourage the XAI, HCXAI, and HCI communities to investigate the real-world settings and design trade-offs in which transparency can be both contextually appropriate (i.e. meaningful according to Cobbe, Norval, and others [69, 247]) and detrimental to non-expert users. To my knowledge, there is no clear indication emerging from the literature as to when *meaningful* transparency becomes detrimental to non-expert users, though this would help to further understand the limitations of transparency in promoting ethical and responsible algorithmic systems for users. In particular the relationship between meaningful transparency and feminist design has not been explored in depth to my knowledge, so Chapter 5 could provide a first step towards it. I would argue this space certainly requires much more academic attention, especially to address the gender data gap that affects many aspects of women’s daily lives, e.g. access to healthcare, safety and security, political power, spending power [55, 54, 276].

The second research avenue that deserves more attention is the challenge of assessing meaningful transparency in practice (Assessment Challenge (RQ3), Table 1.1). Chapters 3 and 6 offer a framework (Nissembaum’s contextual integrity [245]) and a critical method (object-based analysis) towards this, but these have not been tested and adopted by others yet, and I do not claim to have tackled this challenge fully. For example, it is still unclear whether meaningful transparency can be quantified, e.g. with metrics similar to the IEEE 7001 TLs (Chapter 4). I therefore argue further research is needed to use this framework and critical method in other contexts, and other theories could be explored to conceptualise meaningful transparency for non-expert users. Such effort is key to identify best practices and clear examples of systems that are meaningfully transparent for non-expert users, as asked by designers in my study (Chapter 4), and thus move closer to design implementation. These questions should be studied in the context of multi-stakeholder human-computer interactions, to better reflect the real-world setting and challenges, e.g. the ‘many hands’ problem whereby most systems are an ensemble of many systems provided by various deployers [70].

Chapter 8

Conclusion

Algorithmic system transparency is largely needed for non-expert users to understand and challenge a system, among other goals. In this thesis, I have argued that transparency must become *meaningful* to them (i.e. contextually appropriate) to do so, and have explored how to design it in practice. Indeed, without meaningful transparency, it would not be possible to understand, critique, and scrutinise algorithmic systems. My aim is to help bridge the gap between theory and practice towards better aligning such systems with users' information needs. Closing this gap is important for ensuring more responsible algorithmic systems going forward. Towards this, I have identified three main challenges (Table 1.1) to study *how to design meaningful algorithmic system transparency for non-expert users?* (CRQ) :

- Design Challenge (RQ1): it is often unclear for designers how to build transparent algorithmic systems;
- Guideline Challenge (RQ2): recent guidelines and design recommendations must align with designers' perspective and users needs;
- Assessment Challenge (RQ3): we lack a method to assess (meaningful) transparency in practice.

To explore these challenges, I have used a user-centred, empirical, socio-technical, and impact-driven perspective and have highlighted the following four related gaps in the literature: i) the need to better define non-expert users and their specific needs in relation to algorithmic systems (Chapters 3 and 5); ii) the need to explore the gap between new transparency standards and design practices (Chapter 4); iii) the need to better operationalise design recommendations from recent human-centred explainable AI (HCXAI) publications in real-world settings (Chapter 5); and finally iv) the lack of methods to evaluate (meaningful) transparency (Chapters 3 and Chapter 6). I do not claim to have bridged these gaps fully, nor is this list exhaustive. Instead, it provides a starting point to promote meaningful algorithmic systems for non-expert users in practice.

Based on my findings, I conclude that in order to help designers understand how to build transparent algorithmic systems (RQ1), there is a need for a joint effort to raise awareness among users and designers about what algorithmic system transparency can mean in practice (Chapters 3 and 4). Standards promoting transparency can provide a starting point towards this, but their implementation in system design will likely be limited (as they currently are) according to designers (Chapter 4). To answer RQ2, I have therefore shown that direct engagement with designers and users in real-life settings can be productive to better align recent guidelines with designers and users’ perspectives and information needs ((Chapter 5). I have provided methods and illustrative case studies to do so (Chapters 3, 4, and 5). Lastly, to help answer RQ3, I conclude that it is still unclear whether a quantitative measurement of transparency is appropriate or possible in practice (Chapter 4). Instead, I have argued a qualitative assessment method as part of a socio-technical system approach can be applied to help define meaningful transparency for non-expert users in various contexts, with and without direct access to users (Chapters 3 and 6). Overall, I conclude these attempts can inform future work and I have provided several suggestions for further research in the field (Chapter 7).

My main contributions are empirical, critical and theoretical. By exploring the information needs of under-studied groups of non-expert users –i.e. parents in the UK (Chapter 3), gynaecologists in the UK/EU (Chapter 5)– my aim is to provide empirical insights and guidance for designers, academics, and policy-makers on building systems that are meaningfully transparent for different types of non-expert users. By exploring designers’ understanding and experiences making transparent systems in relation to new standards providing transparency requirements (Chapter 4), I contribute towards identifying transparency best practices. I also advocate for designers’ perspectives to be better included when creating new standards in the field. To bridge the ‘principles-to-practice’ gap further, I have tested and expanded design recommendations from the latest HCXAI literature with a co-design approach (Chapter 5). In doing so, I contribute towards complementing and operationalising such recommendations in real-life settings: here a specific clinical practice. To help move towards better assessing meaningful transparency, I have applied Nissembaum’s contextual integrity [245] as a framework to benchmark it in real-life setting (Chapter 3), and have experimented with a critical method, to assess the transparency dimensions of a system through its user interface (Chapter 6). In doing so, I aim to equip designers, regulators, policy-makers, and academics with an accessible, qualitative method that can help them assess transparency gaps in interfaces, so as to better serve users’ information needs. Based on these findings and contributions, I have reached a new conceptualisation of meaningful algorithmic system transparency for non-expert users that can inform future work in the field (Chapter 7).

Finally, I reiterate that transparency is not a panacea and that focusing on non-

expert users can only provide part of the solution towards building more responsible and ethical algorithmic systems for users, hence the need for further research on *meaningful transparency*. I have outlined several research avenues going forward, for example the real-life contexts in which transparency could be contextually appropriate and yet not serve users' needs or interests (see §7.5). Overall, this thesis have both highlighted key challenges and opportunities to promote meaningful transparency for various non-expert users in different settings. I encourage the XAI, HCI, and HCXAI communities to test, expand, and develop these further for and with non-expert users.

Bibliography

- [1] ACM CHI '23. Chi '23 subcommittees, 2022. URL https://chi2023.acm.org/subcommittees/selecting-a-subcommittee/#user_experience.
- [2] Ashraf Abdul, Jo Vermeulen, Danding Wang, Brian Y Lim, and Mohan Kankanhalli. Trends and trajectories for explainable, accountable and intelligible systems: An hci research agenda. In Proceedings of the 2018 CHI conference on human factors in computing systems, pages 1–18, 2018.
- [3] Alessandro Acquisti, Idris Adjerid, and Laura Brandimarte. Gone in 15 seconds: The limits of privacy transparency and control. IEEE Security Privacy, 11(4):72–74, 2013. doi: 10.1109/MSP.2013.86.
- [4] INA Actu. Inscriptions universitaires : l'éternel parcours du combattant. franceinfo ina, 2021. URL https://www.youtube.com/watch?v=AAQN_yZT9nA.
- [5] Rachel Adams. Transparency: New trajectories in law. Routledge, 2020.
- [6] Muhammad Saeed Ahmad, Muhammad Waseem Iqbal, Maham Abid, Nadia Tabasum, Syed Khuram Shahzad, Natash Ali Mian, and Muhammad Raza Naqvi. Usability evaluation of online educational applications in covid-19. Lahore Garrison University Research Journal of Computer Science and Information Technology, 4(4):86–95, 2020.
- [7] The Alan Turing Institute AI Standards Hub. Towards transparent and explainable ai: Workshop to inform iso/iec standards development, 2023. URL <https://aistandardshub.org/events/towards-transparent-and-explainable-ai-workshop-to-inform-iso-iec-standards-d>
- [8] Mamtaj Akter, Amy J Godfrey, Jess Kropczynski, Heather R Lipford, and Pamela J Wisniewski. From parental control to joint family oversight: Can parents and teens manage mobile online safety and privacy as equals? Proceedings of the ACM on Human-Computer Interaction, 6(CSCW1):1–28, 2022.
- [9] Robert H Allen and Ram D Sriram. The role of standards in innovation. Technological Forecasting and Social Change, 64(2-3):171–181, 2000.

- [10] Johannes Allgaier, Lena Mulansky, Rachel Lea Draelos, and Rüdiger Pryss. How does the model make predictions? a systematic literature review on the explainability power of machine learning in healthcare. Artificial Intelligence in Medicine, 143: 102616, 2023.
- [11] Teresa Almeida, Madeline Balaam, Shaowen Bardzell, and Lone Koefoed Hansen. Introduction to the special issue on hci and the body: Reimagining women’s health. ACM Trans. Comput.-Hum. Interact., 27(4), aug 2020. ISSN 1073-0516. doi: 10.1145/3406091. URL <https://doi.org/10.1145/3406091>.
- [12] Mike Ananny and Kate Crawford. Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. New Media & Society, 2016.
- [13] Mike Ananny and Kate Crawford. Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. New Media & Society, 20(3):973–989, 2018.
- [14] Tariq Osman Andersen, Francisco Nunes, Lauren Wilcox, Enrico Coiera, and Yvonne Rogers. Introduction to the special issue on human-centred ai in healthcare: Challenges appearing in the wild, 2023.
- [15] Gloria Andrada, Robert W Clowes, and Paul R Smart. Varieties of transparency: Exploring agency within ai systems. AI & society, 38(4):1321–1331, 2023.
- [16] Anonymous. Ce que contient le programme du nouveau front populaire pour les élections législatives, 14 June 2024. URL https://www.lemonde.fr/politique/article/2024/06/14/ce-que-contient-le-programme-du-nouveau-front-populaire-pour-les-elections-legisla-6239928_823448.html.
- [17] Anna Markella Antoniadis, Yuhang Du, Yasmine Guendouz, Lan Wei, Claudia Mazo, Brett A Becker, and Catherine Mooney. Current challenges and future opportunities for xai in machine learning-based clinical decision support systems: a systematic review. Applied Sciences, 11(11):5088, 2021.
- [18] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bernetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Richard Benjamins, Raja Chatlia, and Francisco Herrera. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. Information Fusion, 58, 2020.

- [19] Vijay Arya, Rachel KE Bellamy, Pin-Yu Chen, Amit Dhurandhar, Michael Hind, Samuel C Hoffman, Stephanie Houde, Q Vera Liao, Ronny Luss, Aleksandra Mojsilović, et al. One explanation does not fit all: A toolkit and taxonomy of ai explainability techniques. arXiv preprint arXiv:1909.03012, 2019.
- [20] John Langshaw Austin. How to Do Things with Words. Oxford University Press, 1975.
- [21] Natalie Avella. Graphic Japan: From Woodblock and Zen to Manga and Kawaii. 2004.
- [22] Karla Badillo-Urquiola, Chhaya Chouhan, Stevie Chancellor, Munmun De Choudhary, and Pamela Wisniewski. Beyond parental control: designing adolescent online safety apps using value sensitive design. Journal of adolescent research, 35(1): 147–175, 2020.
- [23] Carolyn Ball. What is transparency? Public Integrity, 11(4):293–308, 2009.
- [24] Jeffrey Bardzell. Interaction Criticism and Aesthetics, page 2357–2366. Association for Computing Machinery, 2009.
- [25] Jeffrey Bardzell and Shaowen Bardzell. Interaction criticism: a proposal and framework for a new discipline of hci. In CHI’08 Extended Abstracts on Human Factors in Computing Systems, pages 2463–2472. 2008.
- [26] Shaowen Bardzell. Feminist hci: taking stock and outlining an agenda for design. In Proceedings of the SIGCHI conference on human factors in computing systems, pages 1301–1310, 2010.
- [27] Shaowen Bardzell and Jeffrey Bardzell. Towards a feminist hci methodology: social science, feminism, and hci. In Proceedings of the SIGCHI conference on human factors in computing systems, pages 675–684, 2011.
- [28] Ivana Bartoletti. Ai in healthcare: Ethical and privacy challenges. In Artificial Intelligence in Medicine: 17th Conference on Artificial Intelligence in Medicine, AIME 2019, Poznan, Poland, June 26–29, 2019, Proceedings 17, pages 7–10. Springer, 2019.
- [29] Bettina Berendt. The ai act proposal: Towards the next transparency fallacy? why ai regulation should be based on principles based on how algorithmic discrimination works, 2022.
- [30] Marc Berg. Patient care information systems and health care work: a sociotechnical approach. International journal of medical informatics, 55(2):87–101, 1999.

- [31] Astrid Bertrand, Tiphaine Viard, Rafik Belloum, James R Eagan, and Winston Maxwell. On selective, mutable and dialogic xai: a review of what users say about different types of interactive explanations. In Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, pages 1–21, 2023.
- [32] Laurent Bertrandias, Yohan Bernard, and Leila Elgaaied-Gambier. How using parental control software can enhance parents’ well-being: The role of product features on parental efficacy and stress. Journal of Interactive Marketing, 58(2-3): 280–300, 2023.
- [33] Paul Best, Roger Manktelow, and Brian Taylor. Online communication, social media and adolescent wellbeing: A systematic narrative review. Children and Youth Services Review, 41:27–36, 2014.
- [34] Ine Beyens, Eline Frison, and Steven Eggermont. “i don’t want to miss a thing”: Adolescents’ fear of missing out and its relationship to adolescents’ social needs, facebook use, and facebook related stress. Computers in Human Behavior, 64:1–8, 2016.
- [35] Umang Bhatt, Alice Xiang, Shubham Sharma, Adrian Weller, Ankur Taly, Yunhan Jia, Joydeep Ghosh, Ruchir Puri, José M. F. Moura, and Peter Eckersley. Explainable machine learning in deployment. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, FAT* ’20, page 648?657, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450369367. doi: 10.1145/3351095.3375624. URL <https://doi.org/10.1145/3351095.3375624>.
- [36] Umang Bhatt, Javier Antorán, Yunfeng Zhang, Q Vera Liao, Prasanna Sattigeri, Riccardo Fogliato, Gabrielle Melançon, Ranganath Krishnan, Jason Stanley, Omesh Tickoo, et al. Uncertainty as a form of transparency: Measuring, communicating, and using uncertainty. In Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, pages 401–413, 2021.
- [37] Abeba Birhane, William Isaac, Vinodkumar Prabhakaran, Mark Diaz, Madeleine Clare Elish, Iason Gabriel, and Shakir Mohamed. Power to the people? opportunities and challenges for participatory ai. In Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization, EAAMO ’22, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450394772. doi: 10.1145/3551624.3555290. URL <https://doi.org/10.1145/3551624.3555290>.
- [38] Alan F Blackwell. Visual representation. In Mads Soegaard and Rikke Friis Dam,

editors, The encyclopedia of human-computer interaction, chapter 5. The Interaction Design Foundation, 2 edition, 2017.

- [39] Lindsay Blackwell, Emma Gardiner, and Sarita Schoenebeck. Managing expectations: Technology tensions among parents and teens. In Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing, pages 1390–1401, 2016.
- [40] Quentin Blanc. Bientôt des "quotas" dans les classes prépas, 2013. URL <https://etudiant.lefigaro.fr/orientation/actus-et-conseils/detail/article/bientot-des-quotas-dans-les-classes-prepas-1942/>.
- [41] Pauline Bluteau. Le manque de transparence de parcoursup, une fois de plus pointé du doigt, 2020. URL <https://www.letudiant.fr/educpros/actualite/le-manque-de-transparence-de-parcoursup-une-fois-de-plus-pointe-du-doigt.html>.
- [42] Mark Bovens. Analysing and assessing public accountability. a conceptual framework. EUROGOV, European Governance Papers, No. C-06-01, 2006.
- [43] Mark Bovens. Two concepts of accountability: Accountability as a virtue and as a mechanism. In Accountability and European governance, pages 18–39. Routledge, 2014.
- [44] Virginia Braun and Victoria Clarke. Using thematic analysis in psychology. QRP, 3(2):77–101, 2006.
- [45] Virginia Braun and Victoria Clarke. Using thematic analysis in psychology. Qualitative research in psychology, 3(2):77–101, 2006.
- [46] Constance R. Daley Bridges. Schools' Use of Technology in the 21st Century: A Public Relations Concern Regarding the Protection of Children Online. PhD thesis, USA, 2004. AAI3146735.
- [47] John Brooke. SUS: A quick and dirty usability scale. In Usability evaluation in industry. Taylor and Francis, 1996.
- [48] Thomas Buddenkotte, Leonardo Rundo, Ramona Woitek, Lorena Escudero Sanchez, Lucian Beer, Mireia Crispin-Ortuzar, Christian Etmann, Subhadip Mukherjee, Vlad Bura, Cathal McCague, et al. Deep learning-based segmentation of multi-site disease in ovarian cancer. medRxiv, pages 2023–01, 2023.
- [49] Fanny Bugeja-Bloch and Marie-Paule Couto. Le parcoursup des filles. 2018.

- [50] Eleanor R Burgess, Ivana Jankovic, Melissa Austin, Nancy Cai, Adela Kapuścińska, Suzanne Currie, J Marc Overhage, Erika S Poole, and Jofish Kaye. Healthcare ai treatment decision support: Design principles to enhance clinician adoption and trust. In Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, pages 1–19, 2023.
- [51] Alyxander Burns, Christiana Lee, Ria Chawla, Evan Peck, and Narges Mahyar. Who do we mean when we talk about visualization novices? In Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, pages 1–16, 2023.
- [52] Jenna Burrell. How the machine ‘thinks’: Understanding opacity in machine learning algorithms. Big data & society, 3(1):2053951715622512, 2016.
- [53] Adrian Bussone, Simone Stumpf, and Dympna O’Sullivan. The role of explanations on trust and reliance in clinical decision support systems. In 2015 International Conference on Healthcare Informatics, pages 160–169, 2015. doi: 10.1109/ICHI.2015.26.
- [54] Mayra Buvinic and Ruth Levine. Closing the gender data gap. Significance, 13(2): 34–37, 2016.
- [55] Mayra Buvinic, Rebecca Furst-Nichols, and Gayatri Koolwal. Mapping gender data gaps. Data2X report, United Nations Foundation, Washington, DC. http://data2x.org/wp-content/uploads/2014/11/Data2X_MappingGenderDataGaps_FullReport.pdf, 2014.
- [56] BVA. Les adolescents et les smartphones - sondage bva pour wiko, 2018. URL <https://www.bva-group.com/sondages/salaries-mobilite-sondage-bva-salesforce-presse-regionale-2/>.
- [57] Carrie J Cai, Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry. ” hello ai”: uncovering the onboarding needs of medical practitioners for human-ai collaborative decision-making. Proceedings of the ACM on Human-computer Interaction, 3(CSCW):1–24, 2019.
- [58] Rafael A Calvo, Dorian Peters, and Stephen Cave. Advancing impact assessment for intelligent systems. Nature Machine Intelligence, 2(2):89–91, 2020.
- [59] Rafael A Calvo, Dorian Peters, Karina Vold, and Richard M Ryan. Supporting human autonomy in ai systems: A framework for ethical enquiry. In Ethics of Digital Well-Being, pages 31–54. Springer, 2020.

- [60] Tara Capel and Margot Brereton. What is human-centered about human-centered ai? a map of the research landscape. In Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, CHI '23, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450394215. doi: 10.1145/3544548.3580959. URL <https://doi.org/10.1145/3544548.3580959>.
- [61] Linda Charmaraman and Catherine Grevet Delcourt. Prototyping for social well-being with early social media users: Belonging, experimentation, and self-care. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, pages 1–15, 2021.
- [62] Linda Charmaraman, Tracy Gladstone, and Amanda Richer. Positive and negative associations between adolescent mental health and technology. In Technology and adolescent mental health, pages 61–71. Springer, 2018.
- [63] Raja Chatila and John C Havens. The iee global initiative on ethics of autonomous and intelligent systems. In Robotics and well-being, pages 11–16. Springer, 2019.
- [64] Larissa Chazette and Kurt Schneider. Explainability as a non-functional requirement: challenges and recommendations. Requirements Engineering, 25(4):493–514, 2020.
- [65] Lynn Schofield Clark. Parental mediation theory for the digital age. Communication theory, 21(4):323–343, 2011.
- [66] Jennifer Cobbe, Chris Norval, and Jatinder Singh. What lies beneath: Transparency in online service supply chains. Journal of Cyber Policy, 5(1):65–93, 2020.
- [67] Jennifer Cobbe, Michelle Seng Ah Lee, Heleen Janssen, and Jatinder Singh. Centring the rule of law in the digital state. Computer, 53, 2020.
- [68] Jennifer Cobbe, Michelle Seng Ah Lee, and Jatinder Singh. Reviewable automated decision-making: A framework for accountable algorithmic systems. In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, FAccT '21, page 598–609, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383097. doi: 10.1145/3442188.3445921. URL <https://doi.org/10.1145/3442188.3445921>.
- [69] Jennifer Cobbe, Michelle Seng Ah Lee, and Jatinder Singh. Reviewable automated decision-making: A framework for accountable algorithmic systems. In Proceedings of the Conference on Fairness, Accountability, and Transparency, FAccT '21, pages 1–12, New York, USA, 2021. ACM. doi: 10.1145/3442188.3445921.

- [70] Jennifer Cobbe, Michael Veale, and Jatinder Singh. Understanding accountability in algorithmic supply chains. In Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, pages 1186–1197, 2023.
- [71] Collaborative. Wikipedia: Cambridgeshire. <https://en.wikipedia.org/wiki/Cambridgeshire>, .
- [72] Collaborative. Wikipedia: Shropshire. <https://en.wikipedia.org/wiki/Shropshire>, .
- [73] collective. Wikipedia: Ct scans. URL https://en.wikipedia.org/wiki/CT_scan.
- [74] Anne Collier. Less parental control, more support of kids’ self-regulation: Study, 2017.
- [75] European Commision. Proposal for a regulation of the european parliament and of the council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts, 2021. URL <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206>.
- [76] European Commission. Guidelines on automated individual decision-making and profiling for the purposes of regulation 2016/679. https://ec.europa.eu/newsroom/article29/item-detail.cfm?item_id=612053, 2023.
- [77] Commission Nationale de l’Informatique et des Libertés (CNIL). Décision 2017-053 du 30 août 2017, 2017. URL <https://www.legifrance.gouv.fr/cnil/id/CNILTEXT000035647959/>.
- [78] Commission Nationale de l’Informatique et des Libertés (CNIL). Délibération de la formation restreinte no san-2020-012 du 7 décembre 2020 concernant les sociétés google llc et google ireland limited, 2020. URL <https://www.legifrance.gouv.fr/cnil/id/CNILTEXT000042635706>.
- [79] Eric Corbett and Emily Denton. Interrogating the t in facct. In Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, pages 1624–1634, 2023.
- [80] Cours des Comptes. Apb et accès à l’enseignement supérieur : un dispositif contesté à réformer, 2017. URL <https://www.ccomptes.fr/fr/publications/apb-et-acces-lenseignement-superieur-un-dispositif-conteste-reformer>.
- [81] Henriette Cramer, Vanessa Evers, Satyan Ramlal, Maarten Van Someren, Lloyd Rutledge, Natalia Stash, Lora Aroyo, and Bob Wielinga. The effects of transparency

- on trust in and acceptance of a content-based art recommender. User Modeling and User-adapted interaction, 18:455–496, 2008.
- [82] Lorrie Faith Cranor, Adam L Durity, Abigail Marsh, and Blase Ur. {Parents’} and {Teens’} perspectives on privacy in a {Technology-Filled} world. In 10th Symposium On Usable Privacy and Security (SOUPS 2014), pages 19–35, 2014.
 - [83] Karen Croxson, Philippe Bracke, and Carsten Jung. Explaining why computer says no, 31 May 2019. URL <https://webarchive.nationalarchives.gov.uk/ukgwa/20240307162045/https://www.fca.org.uk/insight/explaining-why-computer-says-no>.
 - [84] John Danaher. Three types of algorithmic opacity. Algocracy and the Transhumanist Project, March 2016. URL <https://algocracy.wordpress.com/2016/03/05/three-types-of-algorithmic-opacity>.
 - [85] John Danesh, Yoav Ben-Shlomo, Simon Gault, Jo Semmence, Paul Appleby, Richard Peto, and George Davey Smith. Postcodes as useful markers of social class: population based study in 26 000 british householdscommentary: Socioeconomic position should be measured accurately. Bmj, 318(7187):843–845, 1999.
 - [86] Sir John Daniel. Education and the covid-19 pandemic. Prospects, 49(1):91–96, 2020.
 - [87] Paul B De Laat. Algorithmic decision-making based on machine learning from big data: can transparency restore accountability? Philosophy & technology, 31(4): 525–541, 2018.
 - [88] Ministère de l’Enseignement Supérieur et de la Recherche. Dossier de presse. mise en place du comité scientifique et Éthique, 2018. URL https://cache.media.enseignementsup-recherche.gouv.fr/file/Parcoursup/20/4/DP_PARCOURSUP_COMITE_SCIENTIFIQUE_ET_ETHIQUE_894204.pdf.
 - [89] Eugène Delacroix. Liberty leading the people, 1830. URL https://commons.wikimedia.org/wiki/File:Eug%C3%A8ne_Delacroix_-_Le_28_Juillet._La_Libert%C3%A9_guidant_le_peuple.jpg.
 - [90] Severine Deneulin and Lila Shahani. An introduction to the human development and capability approach: Freedom and agency. IDRC, 2009.
 - [91] Department for Digital, Culture, Media & Sport, Home Office. On-line harms white paper. <https://www.gov.uk/government/consultations/online-harms-white-paper/online-harms-white-paper>, 2020.

- [92] Cour des Comptes. Accès à l'enseignement supérieur : premier bilan de la loi orientation et réussite des étudiants, 2020. URL <https://www.ccomptes.fr/fr/publications/acces-lenseignement-superieur-premier-bilan-de-la-loi-orientation-et-reussite-des>.
- [93] Nicholas Diakopoulos. Accountability, transparency, and algorithms. The Oxford handbook of ethics of AI, 17(4):197, 2020.
- [94] Central Digital and Data Office. Algorithmic transparency standard, 2021. URL <https://www.gov.uk/government/collections/algorithmic-transparency-standard>.
- [95] Dito. Google account dashboard and your data, 2023. URL <https://www.ditoweb.com/2013/09/google-dashboard/#:~:text=Use%20the%20Google%20Dashboard%20to,Youtube%20Videos%2C%20and%20many%20more!>
- [96] Roel Dobbe and Anouk Wolters. Toward sociotechnical ai: Mapping vulnerabilities for machine learning in context. Minds and Machines, 34(2):1–51, 2024.
- [97] Marian Dörk, Patrick Feng, Christopher Collins, and Sheelagh Carpendale. Critical infovis: Exploring the politics of visualization. In CHI '13 Extended Abstracts on Human Factors in Computing Systems, CHI EA '13, page 2189–2198. Association for Computing Machinery, 2013.
- [98] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608, 2017.
- [99] Eleanor Drage, Kerry McInerney, and Jude Browne. Engineers on responsibility: feminist approaches to who's responsible for ethical ai. Ethics and Information Technology, 26(1):4, 2024.
- [100] L Drukker, JA Noble, and AT Papageorghiou. Introduction to artificial intelligence in ultrasound imaging in obstetrics and gynecology. Ultrasound in Obstetrics & Gynecology, 56(4):498–505, 2020.
- [101] Leila H Eadie, Paul Taylor, and Adam P Gibson. A systematic review of computer-assisted diagnosis in diagnostic cancer imaging. European journal of radiology, 81(1):e70–e76, 2012.
- [102] Jacquelynne S Eccles, Christy Miller Buchanan, Constance Flanagan, Andrew Fuligni, Carol Midgley, and Doris Yee. Control versus autonomy during early adolescence. Journal of Social Issues, 47(4):53–68, 1991.

- [103] E&T editorial staff. Truss to ‘tweak’ online safety bill to protect free speech. <https://eandt.theiet.org/content/articles/2022/09/truss-to-tweak-online-safety-bill-to-protect-free-speech/>, 8 September 2022.
- [104] Wikipedia editors. Computer-aided diagnosis. https://en.wikipedia.org/wiki/Computer-aided_diagnosis, 2023.
- [105] Lilian Edwards and Michael Veale. Slave to the algorithm? why a ‘right to an explanation’ is probably not the remedy you are looking for. Duke Law & Technology Review, 17, 2017.
- [106] Lilian Edwards and Michael Veale. Enslaving the algorithm: From a “right to an explanation” to a “right to better decisions”? IEEE Security & Privacy, 16(3): 46–54, 2018.
- [107] Upol Ehsan and Mark O Riedl. Human-centered explainable ai: Towards a reflective sociotechnical approach. In HCI International 2020-Late Breaking Papers: Multimodality and Intelligence: 22nd HCI International Conference, HCII 2020, Copenhagen, Denmark, July 19–24, 2020, Proceedings 22, pages 449–466. Springer, 2020.
- [108] Upol Ehsan, Q. Vera Liao, Michael Muller, Mark O. Riedl, and Justin D. Weisz. Expanding Explainability: Towards Social Transparency in AI Systems. Association for Computing Machinery, 2021. URL <https://doi.org/10.1145/3411764.3445188>.
- [109] Upol Ehsan, Samir Passi, Q Vera Liao, Larry Chan, I Lee, Michael Muller, Mark O Riedl, et al. The who in explainable ai: How ai background shapes perceptions of ai explanations. arXiv preprint arXiv:2107.13509, 2021.
- [110] Upol Ehsan, Philipp Wintersberger, Q. Vera Liao, Martina Mara, Marc Streit, Sandra Wachter, Andreas Riener, and Mark O. Riedl. Operationalizing human-centered perspectives in explainable ai. In Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems, CHI EA ’21, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450380959. doi: 10.1145/3411763.3441342. URL <https://doi.org/10.1145/3411763.3441342>.
- [111] Upol Ehsan, Philipp Wintersberger, Q Vera Liao, Elizabeth Anne Watkins, Carina Manger, Hal Daumé III, Andreas Riener, and Mark O Riedl. Human-centered explainable ai (hcxai): beyond opening the black-box of ai. In CHI conference on human factors in computing systems extended abstracts, pages 1–7, 2022.

- [112] Upol Ehsan, Philipp Wintersberger, Elizabeth A Watkins, Carina Manger, Gonzalo Ramos, Justin D. Weisz, Hal Daumé Iii, Andreas Riener, and Mark O Riedl. Human-centered explainable ai (hcxai): Coming of age. In Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems, CHI EA '23, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450394222. doi: 10.1145/3544549.3573832. URL <https://doi.org/10.1145/3544549.3573832>.
- [113] Malin Eiband, Hanna Schneider, Mark Bilandzic, Julian Fazekas-Con, Mareike Haug, and Heinrich Hussmann. Bringing transparency design into practice. In 23rd international conference on intelligent user interfaces, pages 211–223, 2018.
- [114] Oda Eiichiro. One piece, chapter 1: Romance dawn, 22 July 1997. URL <https://read-onepiecemanga.com/manga/one-piece-chapter-1/>.
- [115] Seymour Epstein. Integration of the cognitive and the psychodynamic unconscious. American psychologist, 49(8):709, 1994.
- [116] Adrian Erasmus, Tyler DP Brunet, and Eyal Fisher. What is interpretability? Philosophy & Technology, 34(4):833–862, 2021.
- [117] Lee B Erickson, Pamela Wisniewski, Heng Xu, John M Carroll, Mary Beth Rosson, and Daniel F Perkins. The boundaries between: Parental involvement in a teen’s online world. Journal of the Association for Information Science and Technology, 67(6):1384–1403, 2016.
- [118] Lorena Escudero Sanchez, Thomas Buddenkotte, Mohammad Al Sa’d, Cathal McCague, James Darcy, Leonardo Rundo, Alex Samoshkin, Martin J Graves, Victoria Hollamby, Paul Browne, et al. Integrating artificial intelligence tools in the clinical research setting: The ovarian cancer use case. Diagnostics, 13(17):2813, 2023.
- [119] Guillemette Faure. ”j’mé suis acheté tout ’stone ocean’ sans me ruiner” : et le pass culture devint le pass mangas, 2021. URL https://www.lemonde.fr/m-le-mag/article/2021/06/03/j-me-suis-achete-tout-stone-ocean-sans-me-ruiner-et-le-pass-culture-devint-le-pass-6082697_4500055.html.
- [120] Heike Felzmann, Eduard Fosch Villaronga, Christoph Lutz, and Aurelia Tamò-Larrieux. Transparency you can trust: Transparency requirements for artificial intelligence between legal norms and contextual concerns. Big Data & Society, 6(1): 2053951719860542, 2019.

- [121] Heike Felzmann, Eduard Fosch-Villaronga, Christoph Lutz, and Aurelia Tamò-Larrieux. Towards transparency by design for artificial intelligence. Science and Engineering Ethics, 26(6):3333–3361, 2020.
- [122] Jessica Fjeld, Nele Achten, Hannah Hilligoss, Adam Nagy, and Madhulika Srikumar. Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for ai. Berkman Klein Center Research Publication, (2020-1), 2020.
- [123] Luciano Floridi, Josh Cows, Monica Beltrametti, Raja Chatila, Patrice Chazerand, Virginia Dignum, Christoph Luetge, Robert Madelin, Ugo Pagallo, Francesca Rossi, et al. Ai4people—an ethical framework for a good ai society: opportunities, risks, principles, and recommendations. Minds and machines, 28:689–707, 2018.
- [124] Mikkel Flyverbom. Transparency: Mediation and the management of visibilities. International Journal of Communication, 10(1):110–122, 2016.
- [125] Office for National Statistics. Exploring the uk’s digital divide. <https://www.ons.gov.uk/peoplepopulationandcommunity/householdcharacteristics/homeinternetandsocialmediausage/articles/exploringtheuksdigitaldivide/2019-03-04>, 4 March 2019.
- [126] Leïla Frouillou, Clément Pin, and Agnès van Zanten. Les plateformes apb et parcoursup au service de l’égalité des chances? L’Annee sociologique, 70(2):337–363, 2020.
- [127] Bright Future. On hack l’algo de parcoursup 2021, 24 January 2021. URL <https://www.youtube.com/watch?v=6rNzdIM07e8>.
- [128] David Gale and Lloyd S Shapley. College admissions and the stability of marriage. The American Mathematical Monthly, 69(1):9–15, 1962.
- [129] Afschin Gandjour. Mutual dependency between capabilities and functionings in amartya sen’s capability approach. Social Choice and Welfare, 31(2):345–350, 2008.
- [130] James Garbarino and K Kostelny. Neighborhood and community influences on parenting. Parenting: An ecological perspective, pages 203–226, 1993.
- [131] Cynthia T Garcia Coll, Elaine C Meyer, and Lisa Brillon. Ethnic and minority parenting. 1995.
- [132] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. Datasheets for datasets. Communications of the ACM, 64(12):86–92, 2021.

- [133] Georgia Wells, Jeff Horwitz and Deepa Seetharaman. The facebook files. <https://www.wsj.com/articles/the-facebook-files-11631713039>, September 2021.
- [134] Anne Gerritsen and Giorgio Riello, editors. Writing material culture history. Writing history. 2021.
- [135] Arup Kumar Ghosh, Karla Badillo-Urquiola, Shion Guha, Joseph J LaViola Jr, and Pamela J Wisniewski. Safety vs. surveillance: what children have to say about mobile apps for parental control. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, pages 1–14, 2018.
- [136] Arup Kumar Ghosh, Karla Badillo-Urquiola, Shion Guha, Joseph J LaViola Jr, and Pamela J Wisniewski. Safety vs. surveillance: what children have to say about mobile apps for parental control. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, pages 1–14, 2018.
- [137] Arup Kumar Ghosh, Karla Badillo-Urquiola, Mary Beth Rosson, Heng Xu, John M Carroll, and Pamela J Wisniewski. A matter of control or safety? examining parental use of technical monitoring apps on teens’ mobile devices. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, pages 1–14, 2018.
- [138] Arup Kumar Ghosh, Charles E Hughes, and Pamela J Wisniewski. Circle of trust: a new approach to mobile online safety for families. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, pages 1–14, 2020.
- [139] Stefani E Gomez. Information practices relative to parental mediation and the family context among Puerto Rican and Dominican teens. Rutgers The State University of New Jersey, School of Graduate Studies, 2017.
- [140] UK Government. Online safety bill. <https://www.gov.uk/society-and-culture/online-safety>, 16 December 2022.
- [141] UK Government. Online safety act press release. <https://bills.parliament.uk/bills/3137/publications>, 26 October 2023.
- [142] Farrah Stone Graham, Susan T Gooden, and Kasey J Martin. Navigating the transparency–privacy paradox in public sector data sharing. The American Review of Public Administration, 46(5):569–591, 2016.
- [143] Séverin Graveleau. Apb 2017 : derniers conseils pour classer ses vœux, 2017. URL https://www.lemonde.fr/bac-lycee/article/2017/05/30/apb-derniers-conseils-pour-classer-ses-v-ux_5136213_4401499.html.

- [144] Ben Green. The contestation of tech ethics: A sociotechnical approach to technology ethics in practice. Journal of Social Computing, 2(3):209–225, 2021.
- [145] Nielson Norman Group. Nielson norman group, 2021. URL <https://www.nngroup.com/>.
- [146] Lara Groves, Jacob Metcalf, Alayna Kennedy, Briana Vecchione, and Andrew Strait. Auditing work: Exploring the new york city algorithmic bias audit regime. In The 2024 ACM Conference on Fairness, Accountability, and Transparency, FAccT '24, page 1107–1120, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400704505. doi: 10.1145/3630106.3658959. URL <https://doi.org/10.1145/3630106.3658959>.
- [147] The Guardian. Online safety bill: changes urged to allow access to social media data. <https://www.theguardian.com/technology/2023/jun/19/online-safety-bill-changes-urged-to-allow-access-to-social-media-data>, 2023.
- [148] Stefano Guerriero, MariaAngela Pascual, Silvia Ajossa, Manuela Neri, Eleonora Musa, Betlem Graupera, Ignacio Rodriguez, and Juan Luis Alcazar. Artificial intelligence (ai) in the detection of rectosigmoid deep endometriosis. European Journal of Obstetrics & Gynecology and Reproductive Biology, 261:29–33, 2021.
- [149] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Gianotti, and Dino Pedreschi. A survey of methods for explaining black box models. ACM computing surveys (CSUR), 51(5):1–42, 2018.
- [150] Angel Gurriá. Openness and transparency - pillars for democracy, trust and progress, 2012. URL <https://www.oecd.org/fr/corruption/opennessandtransparency-pillarsfordemocracytrustandprogress.htm>.
- [151] Atul Guwande. The checklist manifesto. New York: Picadur, 2010.
- [152] AKM Bahalul Haque, AKM Najmul Islam, and Patrick Mikalef. Explainable artificial intelligence (xai) from a user perspective: A synthesis of prior literature and problematizing avenues for future research. Technological Forecasting and Social Change, 186:122120, 2023.
- [153] Peter D Hershock and East-West Center. Ethics and the risks of intelligent technology: the algorithmic threat to freedom of attention. 2020.
- [154] Alexis Hiniker, Hyewon Suh, Sabina Cao, and Julie A Kientz. Screen time tantrums: How families manage screen media experiences for toddlers and preschoolers. In Proceedings of the 2016 CHI conference on human factors in computing systems, pages 648–660, 2016.

- [155] Alexis Hiniker, Jenny S Radesky, Sonia Livingstone, and Alicia Blum-Ross. Moving beyond” the great screen time debate” in the design of technology for children. In Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems, pages 1–6, 2019.
- [156] Wayne Holmes and Ilkka Tuomi. State of the art and practice in ai in education. European Journal of Education, 57(4):542–570, 2022.
- [157] Sarah Homewood. Designing for the changing body: A feminist exploration of self-tracking technologies. In Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems, pages 1–4, 2018.
- [158] Sarah Homewood. Inaction as a design decision: Reflections on not designing self-tracking tools for menopause. In Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems, pages 1–12, 2019.
- [159] Sarah Homewood and Anna Vallgård. Putting phenomenological theories to work in the design of self-tracking technologies. In Proceedings of the 2020 ACM Designing Interactive Systems Conference, pages 1833–1846, 2020.
- [160] Sarah Homewood, Harvey Bewley, and Laurens Boer. Ovum: Designing for fertility tracking as a shared and domestic experience. In Proceedings of the 2019 on Designing Interactive Systems Conference, pages 553–565, 2019.
- [161] Sarah Homewood, Laurens Boer, and Anna Vallgård. Designers in white coats: deploying ovum, a fertility tracking device. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, pages 1–13, 2020.
- [162] Dingdu Hu, Junming Jian, Yongai Li, and Xin Gao. Deep learning-based segmentation of epithelial ovarian cancer on t2-weighted magnetic resonance images. Quantitative Imaging in Medicine and Surgery, 13(3):1464, 2023.
- [163] AI Standards Hub. Ai standards database search, 2022. URL <https://aistandardshub.org/>.
- [164] AI Standards Hub. Ai standards database search, 30 November 2022. URL https://aistandardshub.org/ai-standards-search/?_sft_standards_topic=human-centred-design,human-computer-interaction.
- [165] IBM. Design for ai. explainability, 2019. URL <https://www.ibm.com/design/ai/ethics/explainability>.
- [166] IBM. Transparency and trust in the cognitive era., 20217. URL <https://www.ibm.com/blogs/think/2017/01/ibm-cognitive-principles/>.

- [167] ICO. Key data protection themes. children. <https://ico.org.uk/for-organisations/guide-to-data-protection/key-dp-themes/children/>.
- [168] IEEE. Ieee 7000™-2021. integrating ethical and functional requirements to mitigate risk and increase innovation in systems engineering design and development, 2021. URL <https://engagestandards.ieee.org/ieee-7000-2021-for-systems-design-ethical-concerns.html>.
- [169] IEEE. Ieee, 2021. URL <https://www.ieee.org/>.
- [170] IEEE. P7001/d3 draft standard for transparency of autonomous systems, 2021. URL https://www.techstreet.com/ieee/standards/ieee-p7001?gateway_code=ieee&vendor_id=6929&product_id=2188891.
- [171] IEEE. 7001-2021. ieee standard for transparency of autonomous systems, 2022. URL <https://ieeexplore.ieee.org/document/9726144>.
- [172] Nanna Inie and Peter Dalsgaard. How interaction designers use tools to capture, manage, and collaborate on ideas. In Proceedings of the 2017 ACM Human-Computer Interaction (CHI) Conference Extended Abstracts on Human Factors in Computing Systems, CHI EA '17, page 2668–2675, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450346566. doi: 10.1145/3027063.3053210. URL <https://doi.org/10.1145/3027063.3053210>.
- [173] Institut national de la statistique et des études économiques. 94 % des 15-29 ans ont un smartphone en 2021, 2021. URL <https://www.insee.fr/fr/statistiques/6036909>.
- [174] The British Standards Institute. Iso/iec 30150-1:2022 information technology — affective computing user interface (aui) — part 1: Model, 2022. URL <https://knowledge.bsigroup.com/products/information-technology-affective-computing-user-interface-aui-model/standard/preview>.
- [175] The British Standards Institute. bsigroup.com, 2022. URL <https://www.bsigroup.com/en-GB/standards/>.
- [176] British Standards Institution. Robots and robotic devices: Guide to the ethical design and application of robots and robotic systems, 2016. URL <https://shop.bsigroup.com/products/robots-and-robotic-devices-guide-to-the-ethical-design-and-application-of-robotic-standard>.

- [177] ISO/IEC. Iso/iec awi 12792 information technology — artificial intelligence — transparency taxonomy of ai systems, 2022. URL <https://aistandardshub.org/ai-standards/information-technology-artificial-intelligence-transparency-taxonomy-of-ai-systems>
- [178] Maia Jacobs, Jeffrey He, Melanie F. Pradier, Barbara Lam, Andrew C Ahn, Thomas H McCoy, Roy H Perlis, Finale Doshi-Velez, and Krzysztof Z Gajos. Designing ai for trust and collaboration in time-constrained medical decisions: a sociotechnical lens. In Proceedings of the 2021 chi conference on human factors in computing systems, pages 1–14, 2021.
- [179] Carrie James, Katie Davis, Linda Charmaraman, Sara Konrath, Petr Slovak, Emily Weinstein, and Lana Yarosh. Digital life and youth well-being, social connectedness, empathy, and narcissism. Pediatrics, 140(Supplement_2):S71–S75, 2017.
- [180] Milena Janic, Jan Pieter Wijnbenga, and Thijs Veugen. Transparency enhancing tools (tets): an overview. In 2013 Third Workshop on Socio-Technical Aspects in Security and Trust, pages 18–25. IEEE, 2013.
- [181] Fei Jiang, Yong Jiang, Hui Zhi, Yi Dong, Hao Li, Sufeng Ma, Yilong Wang, Qiang Dong, Haipeng Shen, and Yongjun Wang. Artificial intelligence in healthcare: past, present and future. Stroke and vascular neurology, 2(4), 2017.
- [182] Anna Jobin, Marcello Ienca, and Effy Vayena. The global landscape of ai ethics guidelines. Nature Machine Intelligence, 1(9):389–399, 2019.
- [183] Elena Jost, Philipp Kosian, Jorge Jimenez Cruz, Shadi Albarqouni, Ulrich Gembruch, Brigitte Strizek, and Florian Recker. Evolving the era of 5d ultrasound? a systematic literature review on the applications for artificial intelligence ultrasound imaging in obstetrics and gynecology. Journal of Clinical Medicine, 12(21):6833, 2023.
- [184] Daniel Kahneman. Thinking, fast and slow. macmillan, 2011.
- [185] Dimitra Kamarinou, Christopher Millard, and Jatinder Singh. Machine learning with personal data. In Data protection and privacy: the age of intelligent machines. Hart Publishing, 2017.
- [186] Andreas Kaplan and Michael Haenlein. Siri, siri, in my hand: Who’s the fairest in the land? on the interpretations, illustrations, and implications of artificial intelligence. Business horizons, 62(1):15–25, 2019.
- [187] Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna Wallach, and Jennifer Wortman Vaughan. Interpreting interpretability: understanding data

scientists’ use of interpretability tools for machine learning. In Proceedings of the 2020 CHI conference on human factors in computing systems, pages 1–14, 2020.

- [188] Christopher J Kelly, Alan Karthikesalingam, Mustafa Suleyman, Greg Corrado, and Dominic King. Key challenges for delivering clinical impact with artificial intelligence. BMC medicine, 17:1–9, 2019.
- [189] Os Keyes, Burren Peil, Rua M. Williams, and Katta Spiel. Reimagining (women’s) health: Hci, gender and essentialised embodiment. ACM Trans. Comput.-Hum. Interact., 27(4), aug 2020. ISSN 1073-0516. doi: 10.1145/3404218. URL <https://doi.org/10.1145/3404218>.
- [190] Sara Kiesler, Bozena Zdaniuk, Vicki Lundmark, and Robert Kraut. Troubles with the internet: The dynamics of help at home. Human-computer interaction, 15(4): 323–351, 2000.
- [191] Ansgar Koene, Chris Clifton, Yohko Hatada, Helena Webb, and Rashida Richardson. A governance framework for algorithmic accountability and transparency. 2019.
- [192] Joshua A Kroll. The fallacy of inscrutability. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 376(2133):20180084, 2018.
- [193] Joshua A Kroll. Accountability in computer systems. Oxford University Press New York, 2020.
- [194] Joshua A Kroll. Outlining traceability: A principle for operationalizing accountability in computing systems. In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, pages 758–771, 2021.
- [195] Joshua Alexander Kroll. Accountable algorithms. PhD thesis, Princeton University, 2015.
- [196] Joshua Alexander Kroll, Joanna Huey, Solon Barocas, Edward W. Felten, Joel R. Reidenberg, David G. Robinson, and Harlan Yu. Accountable algorithms. University of Pennsylvania Law Review, 165:633–705, 2017. URL https://scholarship.law.upenn.edu/penn_law_review/vol165/iss3/3/.
- [197] Samuli Laato, Miika Tiainen, AKM Najmul Islam, and Matti Mäntymäki. How to explain ai systems to end users: a systematic literature review and research agenda. Internet Research, 32(7):1–31, 2022.

- [198] Jean-Baptiste Lamy, Boomadevi Sekar, Gilles Guezennec, Jacques Bouaud, and Brigitte Séroussi. Explainable artificial intelligence for breast cancer: A visual case-based reasoning approach. Artificial intelligence in medicine, 94:42–53, 2019.
- [199] Stefan Larsson and Fredrik Heintz. Transparency in artificial intelligence. Internet Policy Review, 9(2), 2020.
- [200] Le Monde. Parcoursup : plus de 90 000 candidats toujours en attente de places, 2021. URL https://www.lemonde.fr/societe/article/2021/07/16/parcoursup-plus-de-91-000-candidats-toujours-en-attente-de-places_6088422_3224.html.
- [201] Soazig Le Nevé. Parcoursup et le mirage des filières ”non sélectives” à l’université, 2 July 2021. URL https://www.lemonde.fr/societe/article/2021/07/02/parcoursup-l-admission-a-l-universite-hypothese-incertaine-pour-les-lyceens_6086625_3224.html.
- [202] Michelle Seng Ah Lee and Jat Singh. The landscape and gaps in open source fairness toolkits. In Proceedings of the 2021 CHI conference on human factors in computing systems, pages 1–13, 2021.
- [203] Michelle Seng Ah Lee, Jennifer Cobbe, Heleen Janssen, and Jatinder Singh. Defining the scope of ai adm system risk assessment. In Research Handbook on EU Data Protection Law, pages 405–434. Edward Elgar Publishing, 2022.
- [204] Seng Ah Lee. Context-conscious fairness throughout the machine learning lifecycle. PhD thesis, University of Cambridge, 2023.
- [205] Natacha Lefauconnier. Apb 2017 : premier tirage au sort en paces !, 2017. URL <https://www.letudiant.fr/educpros/actualite/apb-2017-premier-tirage-au-sort-en-paces.html>.
- [206] Natacha Lefauconnier and Aurore Abdoul-Maninroudine. Apb 2017 : les filières de droit et de staps aux prises avec le tirage au sort, 2017. URL <https://www.letudiant.fr/educpros/actualite/apb-2017-risque-de-tirage-au-sort-en-droit-et-en-staps.html>.
- [207] Legifrance. Loi n°84-52 du 26 janvier 1984 sur l’enseignement supérieur, 1984. URL <https://www.legifrance.gouv.fr/loda/id/JORFTEXT000000692733/>.
- [208] Legifrance. Loi n° 2016-1321 du 7 octobre 2016 pour une république numérique (1), 2016. URL <https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000033202746/>.

- [209] Legifrance. Rapport relatif à la mise en application de la loi n°2016-1321 du 7 octobre 2016 pour une république numérique, 2017. URL https://www.legifrance.gouv.fr/contenu/Media/Files/autour-de-la-loi/legislatif-et-reglementaire/application-des-lois/rapports-art-67/2016/rapport_art67_loi_2016_1321.pdf.
- [210] Legifrance. Arrêté du 9 mars 2018 relatif aux missions, à la composition et aux modalités de fonctionnement du comité éthique et scientifique de la plateforme parcoursup, 2018. URL https://www.google.com/url?q=https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000036692591?r%3D1US6ZPflvF&sa=D&source=editors&ust=1631229612541000&usg=AOvVaw2v8oofQzMq1W3lD0f0PB_a.
- [211] David Lehr and Paul Ohm. Playing with the data: What legal scholars should learn about machine learning. UC Davis Law Review, 51, 2017.
- [212] Damien Leloup. Rapport sur les écrans : les introuvables réseaux sociaux éthiques , 2024. URL https://www.lemonde.fr/pixels/article/2024/05/01/rapport-sur-les-ecrans-les-introuvables-reseaux-sociaux-ethiques_6231000_4408996.html.
- [213] Q Vera Liao, Daniel Gruen, and Sarah Miller. Questioning the ai: informing design practices for explainable ai user experiences. In Proceedings of the 2020 CHI conference on human factors in computing systems, pages 1–15, 2020.
- [214] Brian Y Lim and Anind K Dey. Assessing demand for intelligibility in context-aware applications. In Proceedings of the 11th international conference on Ubiquitous computing, pages 195–204, 2009.
- [215] Brian Y Lim, Anind K Dey, and Daniel Avrahami. Why and why not explanations improve the intelligibility of context-aware intelligent systems. In Proceedings of the SIGCHI conference on human factors in computing systems, pages 2119–2128, 2009.
- [216] Sonia Livingstone and Ellen J Helsper. Parental mediation of children’s internet use. Journal of broadcasting & electronic media, 52(4):581–599, 2008.
- [217] Alex John London. Artificial intelligence and black-box medical decisions: accuracy versus explainability. Hastings Center Report, 49(1):15–21, 2019.
- [218] David Louapre. Parcoursup et les algorithmes de mariage stable, 2020. URL <https://www.youtube.com/watch?v=qsPNqp6555k>.

- [219] David Louapre. Parcoursup et les algorithmes de mariage stable, 2020. URL <https://scienceetonnante.com/2020/01/09/parcoursup/>.
- [220] Jamie Luguri and Lior Jacob Strahilevitz. Shining a light on dark patterns. Journal of Legal Analysis, 13(1):43–109, 2021.
- [221] Michael A Madaio, Luke Stark, Jennifer Wortman Vaughan, and Hanna Wallach. Co-designing checklists to understand organizational challenges and opportunities around fairness in ai. In Proceedings of the 2020 ACM Human-Computer Interaction (CHI) conference on human factors in computing systems, CHI’20, pages 1–14, New York, NY, USA, 2020. Association for Computing Machinery.
- [222] Mahbub ul Haq. Human development index, 1990. URL https://en.wikipedia.org/wiki/Human_Development_Index.
- [223] Gabriel Maicas, Mathew Leonardi, Jodie Avery, Catrina Panuccio, Gustavo Carneiro, M Louise Hull, and George Condous. Deep learning to diagnose pouch of douglas obliteration with ultrasound sliding sign. Reproduction and Fertility, 2(4):236–243, 2021.
- [224] Gianclaudio Malgieri and Frank A Pasquale. From transparency to justification: toward ex ante accountability for ai. Brooklyn Law School, Legal Studies Paper, (712), 2022.
- [225] Anne-Marie Mann, Uta Hinrichs, Janet C. Read, and Aaron Quigley. Facilitator, functionary, friend or foe? studying the role of ipads within learning activities across a school year. In Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, CHI ’16, page 1833–1845, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450333627. doi: 10.1145/2858036.2858251. URL <https://doi.org/10.1145/2858036.2858251>.
- [226] Hiroaki Masaki, Kengo Shibata, Shui Hoshino, Takahiro Ishihama, Nagayuki Saito, and Koji Yatani. Exploring nudge designs to help adolescent sns users avoid privacy and safety threats. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, pages 1–11, 2020.
- [227] David Mascharka, Philip Tran, Ryan Soklaski, and Arjun Majumdar. Transparency by design: Closing the gap between performance and interpretability in visual reasoning. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 4942–4950, 2018.
- [228] Arunesh Mathur, Gunes Acar, Michael J Friedman, Eli Lucherini, Jonathan Mayer, Marshini Chetty, and Arvind Narayanan. Dark patterns at scale: Findings from

a crawl of 11k shopping websites. Proceedings of the ACM on human-computer interaction, 3(CSCW):1–32, 2019.

- [229] Praful Mathur, Karan Kakwani, Srinivas Kudavelly, GA Ramaraju, et al. Deep learning based quantification of ovary and follicles using 3d transvaginal ultrasound in assisted reproduction. In 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), pages 2109–2112. IEEE, 2020.
- [230] Brenna McNally, Priya Kumar, Chelsea Hordatt, Matthew Louis Mauriello, Shalmali Naik, Leyla Norooz, Alazandra Shorter, Evan Golub, and Allison Druin. Co-designing mobile online safety applications with children. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, pages 1–9, 2018.
- [231] Albert Meijer. Transparency. 2014.
- [232] Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. Artificial intelligence, 267:1–38, 2019.
- [233] Ministère de l’Enseignement Supérieur et de la Recherche. Tableaux de bord 2019 des indicateurs de parcoursup, 2019. URL <https://www.enseignementsup-recherche.gouv.fr/cid142122/tableau-de-bord-2019-des-indicateurs-de-parcoursup.html>.
- [234] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. In Proceedings of the conference on fairness, accountability, and transparency, pages 220–229, 2019.
- [235] Kathryn L Modecki, Rachel E Goldberg, Pamela Wisniewski, and Amy Orben. What is digital parenting? a systematic review of past measurement and blueprint for the future. Perspectives on Psychological Science, 17(6):1673–1691, 2022.
- [236] Elhassan Mohamed, Konstantinos Sirlantzis, and Gareth Howells. A review of visualisation-as-explanation techniques for convolutional neural networks and their evaluation. Displays, 73:102239, 2022.
- [237] Le Monde. Parcoursup : des syndicats saisissent le défenseur des droits sur les ”algorithmes locaux”, 2018. URL https://www.lemonde.fr/campus/article/2018/07/17/parcoursup-des-syndicats-saisissent-le-defenseur-des-droits-sur-les-algorithmes-5332773_4401467.html.

- [238] Le Monde. Parcoursup, un système opaque, 2018. URL https://www.lemonde.fr/idees/article/2018/07/18/parcoursup-un-systeme-opaque_5333082_3232.html.
- [239] Alberto Monge Roffarello and Luigi De Russis. The race towards digital wellbeing: Issues and opportunities. In Proceedings of the 2019 CHI conference on human factors in computing systems, pages 1–14, 2019.
- [240] Jessica Morley, Luciano Floridi, Libby Kinsey, and Anat Elhalal. 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, 2020.
- [241] Henrik Mucha, Dennis Mevißen, Sebastian Robert, Ricarda Jacobi, Kirsten Meyer, Winfried Heusler, and Daniel Arzmann. Co-design futures for ai and space: A workbook sprint. In Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems, CHI EA '20, page 1–8, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450368193. doi: 10.1145/3334480.3375203. URL <https://doi.org/10.1145/3334480.3375203>.
- [242] Najme Zehra Naqvi, Kritleen Kaur, Shubhi Khanna, and Shivangi Singh. An overview of machine learning techniques focusing on the diagnosis of endometriosis. Machine Vision and Augmented Intelligence: Select Proceedings of MAI 2022, pages 61–84, 2023.
- [243] The United Nations. The sustainable development goals report 2022. <https://sdgs.un.org/goals>, 2022.
- [244] Sajid Nazir, Diane M Dickson, and Muhammad Usman Akram. Survey of explainable artificial intelligence techniques for biomedical imaging with deep neural networks. Computers in Biology and Medicine, page 106668, 2023.
- [245] Helen Nissenbaum. Privacy as contextual integrity. Wash. L. Rev., 79:119, 2004.
- [246] Helen Nissenbaum. Privacy as contextual integrity. Wash. L. Rev., 79:119, 2004.
- [247] Chris Norval and Jatinder Singh. A room with an overview: Towards meaningful transparency for the consumer internet of things. IEEE Internet of Things Journal, 2023.
- [248] Chris Norval and Jatinder Singh. A room with an overview: Towards meaningful transparency for the consumer internet of things. IEEE Internet of Things Journal, pages 1–1, 2023. doi: 10.1109/JIOT.2023.3318369.

- [249] Chris Norval, Jennifer Cobbe, and Jatinder Singh. Towards an accountable internet of things: A call for ‘reviewability’. In Privacy by Design for the Internet of Things: Building Accountability and Security. The Institution of Engineering and Technology, 2021.
- [250] Chris Norval, Jennifer Cobbe, and Jatinder Singh. Towards an accountable internet of things: A call for ‘reviewability’. In Privacy by Design for the Internet of Things: Building Accountability and Security. The Institution of Engineering and Technology, 2021.
- [251] Chris Norval, Heleen Janssen, Jennifer Cobbe, and Jatinder Singh. Data protection and tech startups: The need for attention, support, and scrutiny. Policy & Internet, 2021.
- [252] Chris Norval, Kristin Cornelius, Jennifer Cobbe, and Jatinder Singh. Disclosure by design: Designing information disclosures to support meaningful transparency and accountability. In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, pages 679–690, 2022.
- [253] Sem Nouws, Marijn Janssen, and Roel Dobbe. Dismantling digital cages: Examining design practices for public algorithmic systems. In International Conference on Electronic Government, pages 307–322. Springer, 2022.
- [254] Martha C Nussbaum and Creating Capabilities. The human development approach. Creating capabilities. Cambridge, MA: Belknap Press of Harvard, 2011.
- [255] Jonathan A. Obar. Sunlight alone is not a disinfectant: Consent and the futility of opening big data black boxes (without assistance). Big Data & Society, 7(1):1–5, 2020.
- [256] OECD Council. Oecd recommendation on public sector information (psi), 2008. URL <https://www.oecd.org/sti/oecdrecommendationonpublicsectorinformationpsi.htm>.
- [257] House of Lords: Select Committee on Artificial Intelligence. Ai in the uk: ready, willing and able? <https://publications.parliament.uk/pa/ld201719/ldselect/ldai/100/100.pdf>, 2018.
- [258] State Internet Information Office of the People’s Republic of China. S.4066 - a bill to require transparency, accountability, and protections for consumers online., 2020. URL http://www.cac.gov.cn/2021-08/27/c_1631652502874117.htm.

- [259] Ofcom UK. Ofcom. making communications work for everyone. <https://www.ofcom.org.uk/>.
- [260] Ofcom Uk. Ofcom to regulate harmful content online. <https://www.ofcom.org.uk/about-ofcom/latest/features-and-news/ofcom-to-regulate-harmful-content-online>, 15 December 2020.
- [261] Ofcom UK, ICO., Jigsaw Research. Internet users’ experience of potential online harms: summary of survey research. https://www.ofcom.org.uk/__data/assets/pdf_file/0024/196413/concerns-and-experiences-online-harms-2020-chart-pack.pdf, 2020.
- [262] Information Commissioner’s Office and The Alan Turing Institute. Explaining decisions made with ai, 2020. URL <https://ico.org.uk/for-organisations/guide-to-data-protection/key-data-protection-themes/explaining-decisions-made-with-ai>.
- [263] Changhoon Oh, Jungwoo Song, Jinhan Choi, Seonghyeon Kim, Sungwoo Lee, and Bongwon Suh. I lead, you help but only with enough details: Understanding user experience of co-creation with artificial intelligence. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, pages 1–13, 2018.
- [264] Gwenn Schurgin O’Keeffe, Kathleen Clarke-Pearson, Council on Communications, and Media. The impact of social media on children, adolescents, and families. Pediatrics, 127(4):800–804, 2011.
- [265] Cathy O’Neil. Weapons of math destruction: How big data increases inequality and threatens democracy. Crown, 2016.
- [266] Nazmun Nisat Ontika, Hussain Abid Syed, Sheree May Saßmannshausen, Richard HR Harper, Yunan Chen, Sun Young Park, Miria Grisot, Astrid Chow, Nils Blaumer, Aparecido Fabiano Pinatti de Carvalho, et al. Exploring human-centered ai in healthcare: diagnosis, explainability, and trust. 2022.
- [267] Amy Orben. Teenagers, screens and social media: a narrative review of reviews and key studies. Social psychiatry and psychiatric epidemiology, 55(4):407–414, 2020.
- [268] Otter.ai. Otter.ai, 2021. URL <https://otter.ai/>.
- [269] Laura M Padilla-Walker, Sarah M Coyne, Ashley M Fraser, W Justin Dyer, and Jeremy B Yorgason. Parents and adolescents growing up in the digital age: Latent growth curve analysis of proactive media monitoring. Journal of adolescence, 35(5): 1153–1165, 2012.

- [270] Google Pair. People + ai guidebook, 2019. URL <https://pair.withgoogle.com/guidebook/chapters>.
- [271] Cecilia Panigutti, Andrea Beretta, Daniele Fadda, Fosca Giannotti, Dino Pedreschi, Alan Perotti, and Salvatore Rinzivillo. Co-design of human-centered, explainable ai for clinical decision support. ACM Transactions on Interactive Intelligent Systems, 2023.
- [272] Erwin Panofsky. Early Netherlandish Painting, Its Origins and Character. Harvard University Press, 1953.
- [273] Nick Park. Wallace and gromit films. <https://www.wallaceandgromit.com/films/>, 2023.
- [274] Frank Pasquale. The Black Box Society: The secret algorithms that control money and information. Harvard University Press, 2015.
- [275] Dino Pedreschi, Fosca Giannotti, Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, and Franco Turini. Meaningful explanations of black box ai decision systems. In Proceedings of the AAAI conference on artificial intelligence, volume 33, pages 9780–9784, 2019.
- [276] Caroline Criado Perez. Invisible women: Data bias in a world designed for men. Abrams, 2019.
- [277] Dorian Peters, Naseem Ahmadpour, and Rafael A Calvo. Tools for wellbeing-supportive design: Features, characteristics, and prototypes. Multimodal Technologies and Interaction, 4(3):40, 2020.
- [278] Thomas Piketty. Le scandale apb, 2016. URL <https://www.lemonde.fr/blog/piketty/2016/07/12/le-scandale-apb/>.
- [279] Hilary Porter. Ecomuseums, 2021. URL <https://ecomuseums.com/>.
- [280] Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Wortman Vaughan, and Hanna Wallach. Manipulating and measuring model interpretability. In Proceedings of the 2021 CHI conference on human factors in computing systems, pages 1–52, 2021.
- [281] Jennifer Preece, Helen Sharp, and Yvonne Rogers. Interaction design: Beyond human-computer interaction. Wiley, 2004.
- [282] Qualtrics. Qualtrics, 2021. URL <https://www.qualtrics.com/fr/?rid=langMatch&prevsite=en&newsite=fr&geo=FR&geomatch=>.

- [283] Jenny Radesky and Alexis Hiniker. From moral panic to systemic change: making child-centered design the default. International Journal of Child-Computer Interaction, 31:100351, 2022.
- [284] Arun Rai. Explainable ai: From black box to glass box. Journal of the Academy of Marketing Science, 48:137–141, 2020.
- [285] Bogdana Rakova, Jingying Yang, Henriette Cramer, and Rumman Chowdhury. Where responsible ai meets reality: Practitioner perspectives on enablers for shifting organizational practices. Proceedings of the ACM on Human-Computer Interaction, 5(CSCW1):1–23, 2021.
- [286] Gonzalo Ramos, Jina Suh, Soroush Ghorashi, Christopher Meek, Richard Banks, Saleema Amershi, Rebecca Fiebrink, Alison Smith-Renner, and Gagan Bansal. Emerging perspectives in human-centered machine learning. In Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems, CHI EA ’19, page 1–8, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450359719. doi: 10.1145/3290607.3299014. URL <https://doi.org/10.1145/3290607.3299014>.
- [287] Guido Rebecchini. What is art history?, 2021. URL <https://courtauld.ac.uk/study/undergraduate/what-is-art-history/>.
- [288] La Nouvelle République.fr. Réforme du bac et parcoursup : les lycéens manifestent ce mardi à niort, 2021. URL <https://www.lanouvellerepublique.fr/niort/reforme-du-bac-et-parcoursup-les-lyceens-manifestent-ce-mardi-a-niort>.
- [289] RhinOcc. Les jeunes et l’inclusion numérique en france : entre exclusion et opportunités., 2020. URL <https://rhinocc.fr/les-jeunes-et-linclusion-numerique-en-france-entre-exclusion-et-opportunites/>.
- [290] Santiago Romero-Brufau, Kirk D Wyatt, Patricia Boyum, Mindy Mickelson, Matthew Moore, and Cheristi Cognetta-Rieke. A lesson in implementation: a pre-post study of providers’ experience with artificial intelligence-based clinical decision support. International journal of medical informatics, 137:104072, 2020.
- [291] Gilad Rosner and Erin Kenneally. Clearly opaque: Privacy risks of the internet of things. In Rosner, Gilad and Kenneally, Erin, Clearly Opaque: Privacy Risks of the Internet of Things (May 1, 2018). IoT Privacy Forum, 2018.
- [292] Mary Beth Rosson, Susanne Maass, and Wendy A. Kellogg. Designing for designers: an analysis of design practice in the real world. SIGCHI Bulletin, 17(SI):137–142,

May 1986. ISSN 0736-6906. doi: 10.1145/30851.30873. URL <https://doi.org/10.1145/30851.30873>.

- [293] Royal anthropological institute, the Department of Africa, Oceania and the Americas of the British Museum and the Department of Anthropology at SOAS. Art, materiality and representation conference, 2018. URL <https://www.therai.org.uk/conferences/art-materiality-and-representation>.
- [294] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature machine intelligence, 1(5): 206–215, 2019.
- [295] Stuart J Russell and Peter Norvig. Artificial intelligence a modern approach. London, 2010.
- [296] Tara L Rutkowski, Heidi Hartikainen, Kirsten E Richards, and Pamela J Wisniewski. Family communication: Examining the differing perceptions of parents and teens regarding online safety communication. Proceedings of the ACM on Human-Computer Interaction, 5(CSCW2):1–23, 2021.
- [297] Matthew N Sadiku, Mahamadou Tembely, and Sarhan M Musa. Digital parenting. Journal of Multidisciplinary Engineering Science and Technology, 4(1):6454–6456, 2017.
- [298] Waddah Saeed and Christian Omlin. Explainable ai (xai): A systematic meta-survey of current challenges and future opportunities. Knowledge-Based Systems, 263: 110273, 2023.
- [299] Elizabeth B-N Sanders and Pieter Jan Stappers. Co-creation and the new landscapes of design. Co-design, 4(1):5–18, 2008.
- [300] Mike Savage, Fiona Devine, Niall Cunningham, Mark Taylor, Yaojun Li, Johs Hjellbrekke, Brigitte Le Roux, Sam Friedman, and Andrew Miles. A new model of social class? findings from the bbc’s great british class survey experiment. Sociology, 47(2):219–250, 2013.
- [301] Diane J Schiano and Christine Burg. Parental controls: Oxymoron and design opportunity. In HCI International 2017–Posters’ Extended Abstracts: 19th International Conference, HCI International 2017, Vancouver, BC, Canada, July 9–14, 2017, Proceedings, Part II 19, pages 645–652. Springer, 2017.
- [302] Diane J Schiano, Christine Burg, Anthony Nalan Smith, and Florencia Moore. Parenting digital youth: How now? In Proceedings of the 2016 CHI Conference

Extended Abstracts on Human Factors in Computing Systems, pages 3181–3189, 2016.

- [303] Daniel Schiff, Bogdana Rakova, Aladdin Ayesh, Anat Fanti, and Michael Lennon. Principles to practices for responsible ai: closing the gap. arXiv preprint arXiv:2006.04707, 2020.
- [304] Bianca G. S. Schor, Chris Norval, Ellen Martin Charlesworth, and Jatinder Singh. Mind the gap: Designers and standards on algorithmic system transparency for users. In Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24). ACM, 2024. URL <https://doi.org/10.1145/3613904.3642531>.
- [305] Bianca Giulia Sarah Schor, Emma Kallina, Jatinder Singh, and Alan Blackwell. Meaningful transparency for clinicians: Operationalising hexai research with gynaecologists. In The 2024 ACM Conference on Fairness, Accountability, and Transparency, FAccT '24, page 1268–1281, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400704505. doi: 10.1145/3630106.3658971. URL <https://doi.org/10.1145/3630106.3658971>.
- [306] Jessica M Schwartz, Amanda J Moy, Sarah C Rossetti, Noémie Elhadad, and Kenrick D Cato. Clinician involvement in research on machine learning-based predictive clinical decision support for the hospital setting: A scoping review. Journal of the American Medical Informatics Association, 28(3):653–663, 2021.
- [307] Amartya Sen. Inequality reexamined. Harvard University Press, 1995.
- [308] Amartya Sen. Development as freedom / Amartya Sen. 2001. ISBN 9780307874290.
- [309] Mark Sendak, Madeleine Clare Elish, Michael Gao, Joseph Futoma, William Ratliff, Marshall Nichols, Armando Bedoya, Suresh Balu, and Cara O'Brien. "the human body is a black box" supporting clinical decision-making with deep learning. In Proceedings of the 2020 conference on fairness, accountability, and transparency, pages 99–109, 2020.
- [310] Oshani Seneviratne and Lalana Kagal. Enabling privacy through transparency. In 2014 Twelfth Annual International Conference on Privacy, Security and Trust, pages 121–128. IEEE, 2014.
- [311] Rushabh Shah and Alina Chircu. Iot and ai in healthcare: A systematic literature review. Issues in Information Systems, 19(3), 2018.
- [312] Ruey-Kai Sheu and Mayuresh Sunil Pardeshi. A survey on medical explainable ai (xai): Recent progress, explainability approach, human interaction and scoring system. Sensors, 22(20):8068, 2022.

- [313] Ronald L Simons, Christine Johnson, Rand D Conger, and Frederick O Lorenz. Linking community context to quality of parenting: A study of rural families 1. Rural Sociology, 62(2):207–230, 1997.
- [314] Jatinder Singh, Jennifer Cobbe, and Chris Norval. Decision provenance: Harnessing data flow for accountable systems. IEEE Access, 7:6562–6574, 2019.
- [315] Venkatesh Sivaraman, Leigh A Bukowski, Joel Levin, Jeremy M Kahn, and Adam Perer. Ignore, trust, or negotiate: understanding clinician acceptance of ai-based treatment recommendations in health care. In Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, pages 1–18, 2023.
- [316] Peter Slattery, Alexander K Saeri, and Peter Bragge. Research co-design in health: a rapid overview of reviews. Health research policy and systems, 18(1):1–13, 2020.
- [317] Helen Smith. Clinical ai: opacity, accountability, responsibility and liability. Ai & Society, 36(2):535–545, 2021.
- [318] Kacper Sokol and Peter Flach. Explainability fact sheets: a framework for systematic assessment of explainable approaches. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, pages 56–67, 2020.
- [319] Anna Spagnolli, Lily E Frank, Pim Haselager, and David Kirsh. Transparency as an ethical safeguard. In Symbiotic Interaction: 6th International Workshop, Symbiotic 2017, Eindhoven, The Netherlands, December 18–19, 2017, Revised Selected Papers 6, pages 1–6. Springer, 2018.
- [320] Bernd Carsten Stahl, Andreas Andreou, Philip Brey, Tally Hatzakis, Alexey Kirichenko, Kevin Macnish, S Laulhé Shaelou, Andrew Patel, Mark Ryan, and David Wright. Artificial intelligence for human flourishing—beyond principles for machine learning. Journal of Business Research, 124:374–388, 2021.
- [321] Marc Steen, Menno Manschot, and Nicole De Koning. Benefits of co-design in service design projects. International Journal of Design, 5(2), 2011.
- [322] Laurence Steinberg. Autonomy, conflict, and harmony in the family relationship. 1990.
- [323] Kristin Stewart, Glen Brodowsky, and Donald Sciglimpaglia. Parental supervision and control of adolescents’ problematic internet use: understanding and predicting adoption of parental control software. Young Consumers, 23(2):213–232, 2022.

- [324] Cynthia Stohl, Michael Stohl, and Paul M Leonardi. Managing opacity: Information visibility and the paradox of transparency in the digital age. IJoC, 10(2016):123–137, 2016.
- [325] Cynthia Stohl, Michael Stohl, and Paul M Leonardi. Digital age— managing opacity: Information visibility and the paradox of transparency in the digital age. International Journal of Communication, 10:15, 2016.
- [326] Marilyn Strathern. The tyranny of transparency. British educational research journal, 26(3):309–321, 2000.
- [327] Kaiwen Sun, Yixin Zou, Jenny Radesky, Christopher Brooks, and Florian Schaub. Child safety in the smart home: Parents’ perceptions, needs, and mitigation strategies. Proceedings of the ACM on Human-Computer Interaction, 5(CSCW2):1–41, 2021.
- [328] Cass R. Sunstein. Why nudge? : the politics of libertarian paternalism. The Storrs Lectures Series. Yale University Press, 2014.
- [329] Pierre-André Taguieff. Macron : miracle ou mirage? Éditions de l’Observatoire, 2017.
- [330] Elda Tartari. Benefits and risks of children and adolescents using social media. European Scientific Journal, 11(13), 2015.
- [331] Shu Wen Tay, Paul Ryan, and C Anthony Ryan. Systems 1 and 2 thinking processes and cognitive reflection testing in medical students. Canadian medical education journal, 7(2):e97, 2016.
- [332] the International Electrotechnical Commission. iec.ch, 2022. URL <https://www.iec.ch/homepage>.
- [333] the International Organization for Standardization. iso.org, 2022. URL <https://www.iso.org/home.html>.
- [334] Anja Thieme, Maryann Hanratty, Maria Lyons, Jorge Palacios, Rita Faia Marques, Cecily Morrison, and Gavin Doherty. Designing human-centered ai for mental health: Developing clinically relevant applications for online cbt treatment. ACM Transactions on Computer-Human Interaction, 30(2):1–50, 2023.
- [335] Vincent Tiberj. Parcoursup ou la sélection par les algorithmes. 2021.
- [336] Sana Tonekaboni, Shalmali Joshi, Melissa D McCradden, and Anna Goldenberg. What clinicians want: contextualizing explainable machine learning for clinical end use. In Machine learning for healthcare conference, pages 359–380. PMLR, 2019.

- [337] Hugo Travers. Parcoursup - ce qu'il faut savoir, 2018. URL <https://www.youtube.com/watch?v=qsPNqp6555k>.
- [338] Edward R Tufte. The visual display of quantitative information. Graphics press Cheshire, CT, 1983.
- [339] Matteo Turilli and Luciano Floridi. The ethics of information transparency. Ethics and Information Technology, 11:105–112, 2009.
- [340] Kazuki Uehara, Masahiro Murakawa, Hirokazu Nosato, and Hidenori Sakanashi. Multi-scale explainable feature learning for pathological image analysis using convolutional neural networks. In 2020 IEEE International Conference on Image Processing (ICIP), pages 1931–1935. IEEE, 2020.
- [341] Yalda T Uhls, Nicole B Ellison, and Kaveri Subrahmanyam. Benefits and costs of social media in adolescence. Pediatrics, 140(Supplement_2):S67–S70, 2017.
- [342] Ofcom UK. Children and parents: Media use and attitudes report 2020/2021. <https://www.ofcom.org.uk/research-and-data/media-literacy-research/childrens/children-and-parents-media-use-and-attitudes-report-2021>, 2021.
- [343] European Union. Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/ec (general data protection regulation). Official Journal of the European Union, L119:1–88, May 2016.
- [344] unknown. askaboutgames. <https://www.askaboutgames.com/>, .
- [345] unknown. commonsensemedia. <https://www.commonsensemedia.org/>, .
- [346] Ibo Van de Poel. Embedding values in artificial intelligence (ai) systems. Minds and machines, 30(3):385–409, 2020.
- [347] Michael Veale, Max Van Kleek, and Reuben Binns. Fairness and accountability design needs for algorithmic support in high-stakes public sector decision-making. CoRR, 2018. URL <http://arxiv.org/abs/1802.01029>.
- [348] Philippe Verduyn, David Seungjae Lee, Jiyoung Park, Holly Shablack, Ariana Orvell, Joseph Bayer, Oscar Ybarra, John Jonides, and Ethan Kross. Passive facebook usage undermines affective well-being: Experimental and longitudinal evidence. Journal of Experimental Psychology: General, 144(2):480, 2015.

- [349] Himanshu Verma, Roger Schaer, Julien Reichenbach, Mario Jreige, John O Prior, Florian Evéquo, and Adrien Depeursinge. On improving physicians’ trust in ai: Qualitative inquiry with imaging experts in the oncological domain. BMC Medical Imaging, in review, 2021.
- [350] Himanshu Verma, Jakub Mlynar, Roger Schaer, Julien Reichenbach, Mario Jreige, John Prior, Florian Evéquo, and Adrien Depeursinge. Rethinking the role of ai with physicians in oncology: Revealing perspectives from clinical and research workflows. In Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, CHI ’23, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450394215. doi: 10.1145/3544548.3581506. URL <https://doi.org/10.1145/3544548.3581506>.
- [351] Sandra Wachter, Brent Mittelstadt, and Chris Russell. Bias preservation in machine learning: the legality of fairness metrics under eu non-discrimination law. W. Va. L. Rev., 123:735, 2020.
- [352] Ben Wagner, Krisztina Rozgonyi, Marie-Therese Sekwenz, Jennifer Cobbe, and Jatinder Singh. Regulating transparency? facebook, twitter and the german network enforcement act. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, pages 261–271, 2020.
- [353] Andrew Walter. Governing finance: East Asia’s adoption of international standards. Cornell University Press, 2011.
- [354] Dakuo Wang, Liuping Wang, Zhan Zhang, Ding Wang, Haiyi Zhu, Yvonne Gao, Xiangmin Fan, and Feng Tian. “brilliant ai doctor” in rural clinics: Challenges in ai-powered clinical decision support system deployment. In Proceedings of the 2021 CHI conference on human factors in computing systems, pages 1–18, 2021.
- [355] Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y Lim. Designing theory-driven user-centric explainable ai. In Proceedings of the 2019 CHI conference on human factors in computing systems, pages 1–15, 2019.
- [356] Liuping Wang, Zhan Zhang, Dakuo Wang, Weidan Cao, Xiaomu Zhou, Ping Zhang, Jianxing Liu, Xiangmin Fan, and Feng Tian. Human-centered design and evaluation of ai-empowered clinical decision support systems: a systematic review. Frontiers in Computer Science, 5:1187299, 2023.
- [357] Qiaosi Wang, Michael Madaio, Shaun Kane, Shivani Kapania, Michael Terry, and Lauren Wilcox. Designing responsible ai: Adaptations of ux practice to meet responsible ai challenges. In Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, pages 1–16, 2023.

- [358] Qiaosi Wang, Michael Madaio, Shaun Kane, Shivani Kapania, Michael Terry, and Lauren Wilcox. Designing responsible ai: adaptations of ux practice to meet responsible ai challenges. In Proceedings of the 2023 ACM Human-Computer Interaction (CHI) Conference on Human Factors in Computing Systems, CHI'23, pages 1–16, New York, NY, USA, 2023. Association for Computing Machinery.
- [359] Adrian Weller. Transparency: motivations and challenges. In Explainable AI: Interpreting, Explaining and Visualizing Deep Learning, pages 23–40. Springer, 2019.
- [360] Maranke Wieringa. What to account for when accounting for algorithms: A systematic literature review on algorithmic accountability. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, FAT* '20, page 1–18, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450369367. doi: 10.1145/3351095.3372833. URL <https://doi.org/10.1145/3351095.3372833>.
- [361] wikipedia. Smartphone. <https://en.wikipedia.org/wiki/Smartphone>.
- [362] Wikipedia. Projet de loi devaquet, 2021. URL https://fr.wikipedia.org/wiki/Projet_de_loi_Devaquet.
- [363] Leland Wilkinson. The grammar of graphics. Springer Science & Business Media, 2013.
- [364] Alan FT Winfield, Serena Booth, Louise A Dennis, Takashi Egawa, Helen Hastie, Naomi Jacobs, Roderick I Muttram, Joanna I Olszewska, Fahimeh Rajabiyazdi, Andreas Theodorou, et al. Ieee p7001: a proposed standard on transparency. Frontiers in Robotics and AI, page 225, 2021.
- [365] Pamela Wisniewski, Heng Xu, Jack Carroll, and Mary Beth Rosson. Grand challenges of researching adolescent online safety: a family systems approach. 2013.
- [366] Pamela J Wisniewski, Jessica Vitak, and Heidi Hartikainen. Privacy in adolescence. In Modern socio-technical perspectives on privacy, pages 315–336. Springer International Publishing Cham, 2022.
- [367] Wix. Wix.com, 2021. URL <https://www.wix.com/>.
- [368] Michel Wlassikoff. Histoire du graphisme en France. Arts décoratifs, 2005.
- [369] Sijia Xiao, Coye Cheshire, and Niloufar Salehi. Sensemaking, support, safety, retribution, transformation: A restorative justice approach to understanding adolescents'

- needs for addressing online harm. In Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems, CHI '22, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450391573. doi: 10.1145/3491102.3517614. URL <https://doi.org/10.1145/3491102.3517614>.
- [370] Yao Xie, Ge Gao, and Xiang 'Anthony' Chen. Outlining the design space of explainable intelligent systems for medical diagnosis. CoRR, abs/1902.06019, 2019. URL <http://arxiv.org/abs/1902.06019>.
- [371] Yao Xie, Melody Chen, David Kao, Ge Gao, and Xiang'Anthony' Chen. Chexplain: enabling physicians to explore and understand data-driven, ai-enabled medical imaging analysis. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, pages 1–13, 2020.
- [372] Feiyu Xu, Hans Uszkoreit, Yangzhou Du, Wei Fan, Dongyan Zhao, and Jun Zhu. Explainable ai: A brief survey on history, research areas, approaches and challenges. In Natural Language Processing and Chinese Computing: 8th CCF International Conference, NLPCC 2019, Dunhuang, China, October 9–14, 2019, Proceedings, Part II 8, pages 563–574. Springer, 2019.
- [373] Qian Xu, Wenzhao Xie, Bolin Liao, Chao Hu, Lu Qin, Zhengzijin Yang, Huan Xiong, Yi Lyu, Yue Zhou, Aijing Luo, et al. Interpretability of clinical decision support systems based on artificial intelligence from technological and medical perspective: A systematic review. Journal of Healthcare Engineering, 2023, 2023.
- [374] Qian Yang, Aaron Steinfeld, and John Zimmerman. 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, pages 1–11, 2019.
- [375] Qian Yang, Aaron Steinfeld, Carolyn Rosé, and John Zimmerman. 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, pages 1–13, 2020.
- [376] Qian Yang, Yuexing Hao, Kexin Quan, Stephen Yang, Yiran Zhao, Volodymyr Kuleshov, and Fei Wang. Harnessing biomedical literature to calibrate clinicians' trust in ai decision support systems. In Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, CHI '23, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450394215. doi: 10.1145/3544548.3581393. URL <https://doi.org/10.1145/3544548.3581393>.

- [377] Sarita Yardi and Amy Bruckman. Social and technical challenges in parenting teens’ social media use. In Proceedings of the SIGCHI conference on human factors in computing systems, pages 3237–3246, 2011.
- [378] Sarita Yardi and Amy Bruckman. Income, race, and class: exploring socioeconomic differences in family technology use. In Proceedings of the SIGCHI conference on human factors in computing systems, pages 3041–3050, 2012.
- [379] Nur Yildirim, Mahima Pushkarna, Nitesh Goyal, Martin Wattenberg, and Fernanda Viégas. 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, pages 1–13, 2023.
- [380] Mireia Yurrita, Tim Draws, Agathe Balayn, Dave Murray-Rust, Nava Tintarev, and Alessandro Bozzon. Disentangling fairness perceptions in algorithmic decision-making: The effects of explanations, human oversight, and contestability. In Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, CHI ’23, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450394215. doi: 10.1145/3544548.3581161. URL <https://doi.org/10.1145/3544548.3581161>.
- [381] Tal Z Zarsky. Transparent predictions. U. Ill. L. Rev., page 1503, 2013.
- [382] John Zerilli, Alistair Knott, James Maclaurin, and Colin Gavaghan. Transparency in algorithmic and human decision-making: is there a double standard? Philosophy & Technology, 32:661–683, 2019.
- [383] Yu Zhang, Peter Tiño, Aleš Leonardis, and Ke Tang. A survey on neural network interpretability. IEEE Transactions on Emerging Topics in Computational Intelligence, 5(5):726–742, 2021.
- [384] Yunfeng Zhang, Q Vera Liao, and Rachel KE Bellamy. Effect of confidence and explanation on accuracy and trust calibration in ai-assisted decision making. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, pages 295–305, 2020.
- [385] Yayan Zhao, Mingwei Li, and Matthew Berger. Graphical perception of saliency-based model explanations. In Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, pages 1–15, 2023.
- [386] Roberto V Zicari, Sheraz Ahmed, Julia Amann, Stephan Alexander Braun, John Brodersen, Frédérick Bruneault, James Brusseau, Erik Campano, Megan Coffee, Andreas Dengel, et al. Co-design of a trustworthy ai system in healthcare: deep learning based skin lesion classifier. Frontiers in Human Dynamics, 3:688152, 2021.

- [387] ZOOM.us. Zoom, 2021. URL <https://zoom.us/>.
- [388] Cui Zou, Wangchuchu Zhao, and Keng Siau. Covid-19 pandemic: A usability study on platforms to support elearning. In International Conference on Human-Computer Interaction, pages 333–340. Springer, 2020.
- [389] Anneke Zuiderwijk, Yu-Che Chen, and Fadi Salem. Implications of the use of artificial intelligence in public governance: A systematic literature review and a research agenda. Government information quarterly, 38(3):101577, 2021.
- [390] Douglas Zytco, Pamela J. Wisniewski, Shion Guha, Eric P. S. Baumer, and Min Kyung Lee. Participatory design of ai systems: Opportunities and challenges across diverse users, relationships, and application domains. In Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems, CHI EA '22, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450391566. doi: 10.1145/3491101.3516506. URL <https://doi.org/10.1145/3491101.3516506>.
- [391] Comité éthique et scientifique de Parcoursup (CESP). Rapport du comité éthique et scientifique de parcoursup, 2019. URL <https://www.enseignementsup-recherche.gouv.fr/cid138009/www.enseignementsup-recherche.gouv.fr/cid138009/rapport-du-comite-ethique-et-scientifique-de-parcoursup.html>.

Appendix A

Chapter 3 Appendix

A.1 Interview script

The following sections provide the script and examples of questions we have asked participants during the online, semi-structured interviews.

A.1.1 Introduction.

Thank you very much for your time, this interview will last about 40 minutes. We will ask you about your experience and concerns as a parent regarding your 10-15yo child(ren) using social media and online games. If you have several children, please bear in mind we are only asking you about those in this age range. We will then ask you to do a short drawing exercise (with pen and paper) to help us improve the design of social media and online games for parents like yourself. There is no right or wrong answer, we are not judging your answers, so feel free to express anything you feel might be relevant. All the questions are optional, so feel free to also skip any question you don't feel like answering or let me know if you wish to stop or withdraw from the study at any point. Is it okay with you?

A.1.2 General experience, concerns, and information needs regarding social media and online games used by children.

- Are you concerned about your 10-15yo using social media and online games? What is your main concern? Why? What social media or online game do you feel poses the most online risk(s) to your 10-15yo? Which risk(s)? Why?
- What social media and online games does your 10-15yo uses the most? How long do they spend using them per day/week? Has your 10-15yo ever been exposed to

an online risk? If so, which one? How do you currently make sure your 10-15yo is not impacted by online harms? Would you like to change the way you are informed about what is going on online when your child uses social media or online games?

- How did you learn about your 10-15yo experiencing this online harm? What information do you think you need about the social media and online games they use in such a situation/ in relation to your concern about...? Do you feel that you received that information? Why/How? How would you like or expect to receive it? How would this information be presented to you?
- When do you think you need that information the most? What follow-up action(s) will be enabled if you receive such information do you think?
- What about the information you would need for this other concern or social media/game you previously mentioned? Do you need the same information? Why? How would you want or expect it to be represented?
- Additional questions: what would make such information more accurate and reliable, and what would you want or find helpful? What kind of information do you think your child(ren) would need about the online games and/or social media they use; what you would like to know, should your child come and ask you for advice/help? Has this happened to you before?

A.1.3 Drawing exercise.

- Based on what we have talked about, I'd like to suggest a short drawing exercise for you now. In the situation where your 10-15yo uses the social media/online game XXX, can you try to draw what an app on a smartphone could look like in order to give you the information that you need about the social media/online game they use? Bear in mind **we are not trying to design a parental control app but rather an online service to give you the information you think might be helpful to understand what is going on behind the screen** when your child uses social media and/or online games.
- You said you would need to know XXX. Can you try to describe orally and visually how that might look like on such an app providing information for parents like yourself?
- Why did you choose this option? Do you feel this drawing you have just made would fully provide you with the information that you need when your child uses this social

media/online game (in the context of your concern about XXX)? Why? What could be missing there in your opinion? Does it feel like this app would be useful to you? Why? Is there any improvement you think could be made? Why and how?

A.1.4 Demographics.

- Age
- Gender
- Highest diploma or degree
- Technical literacy: How confident do you feel using a computer?
- Family situation: living with a long-term partner, single parent, widow(er)...
- Nb of children, their age and gender
- approximate yearly household income

A.1.5 Feedback and snowball sampling.

- Anything you would like to add or have been thinking about during this interview? Do you feel like you have learned something today? If so, what?
- Snowball sampling question: “Are you happy to introduce me to two other parents who might be interested in taking part in this study? Ideally, they would be from two different families in your area and have at least one 10-15yo who uses social media and/or online games.”

Appendix B

Chapter 4 Appendix

B.1 Glossary

In this paper, we use the following definitions from IEEE 7001 [170]:

- **Autonomous system:** a system that has the capacity to make decisions itself in response to some input data or stimulus with a varying degree of human oversight or intervention depending on the system’s level of autonomy.
- **Designers:** designers, developers, builders, maintainers, and operators, as well as decision-makers and procurers in organizations using and deploying autonomous systems (collectively, “designers”) of autonomous systems who either wish to or are required to engineer systems that have a certain degree of transparency.
- **Non-expert users (users):** persons who have only a brief interaction or who interact every day with an autonomous system.
- **Transparency:** a transfer of information from an autonomous system or its designers to a stakeholder that is truthful; contains information relevant to the causes of some action, decision, or behavior; and is presented at a level of abstraction and in a form meaningful to the stakeholder. Transparency should be mindful of the stakeholders’ likely perception and comprehension, and should avoid disclosing information in a manner that, while technically true, is framed in a way that leads to misapprehension.
- **(IEEE 7001) Transparency level for users (TL):** “a requirement, expressed as a qualitative property of the system which must be met” in terms of transparency for users
- **TL0:** “barely or no transparency”
- **TL1:** “accessible information including scenarios and general principles of operation”

- **TL2:** “interactive training material”
- **TL3:** “a functionality to get a brief and immediate explanation of the system’s most recent activity”
- **TL4:** “a functionality to get a brief and immediate explanation of the system’s activity in a given situation”
- **TL5:** “a continuous explanation of the behaviour, which adapts to the user’s needs and context”

B.2 Scenarios

In Study 2 (interviews), participants were presented on slides one of the following scenarios, extracted from IEEE 7001 [170]. Note that we did not change the wording, including the American spelling, except for the word “characteristics,” which refers here to the transparency levels TL1, 2 and 3, as defined in IEEE 7001, so as not to bias our study in case participants were already familiar with IEEE 7001. Due to lack of time, we presented each participant with only one TL and one scenario in total.

B.2.1 *Moderation:* Content Moderation for AI.

- We use this fictional scenario to explore how to design 3 characteristics to provide system transparency for non-expert end-users.
- A video hosting website has been accused by activists of using keywords to prevent monetization of potentially controversial content. To mitigate a potential scandal, the website decides to communicate more transparently the decision-making processes to content creators.
- The content creators are the non-expert end-users here. They require a medium level of understanding of how the system functions, including the ability to ask the system to explain its decisions, or to pre-emptively interrogate if something is likely to be deemed problematic.

B.2.2 Scenario *Credit:* Credit Scoring System.

- We use this fictional scenario to explore how to design 3 characteristics to provide system transparency for non-expert end-users.

- A loans company asks their credit scoring provider to apply 3 characteristics to their technology, to more transparently communicate the decision-making processes to loan applicants.
- The loan applicants are the non-expert end-users here. Transparency is very important to them as the assessment is of their own particulars, and they deserve a chance to understand why they have been assessed in a particular way, and to seek to redress in the event that information is incorrect or is assessed unfairly.

B.3 Summary of Demographics from our survey (Study 1)

Figure B.1 summarises the demographic information provided by our n=40 surveyed designers (Study 1). The responses to the categories ‘self-described profession,’ ‘experience,’ ‘sector,’ and ‘area’ are independent from each other.

Self-described profession	Designer	Developer	Engineer/Technical Manager	Academic	Product/Business Manager	Entrepreneur
Total respondents n=39	7 (18%)	8 (20.5%)	6 (15%)	5 (13%)	8 (20.5%)	5 (13%)
Experience	1 year	1-3 years	4-6 years	7-9 years	10+ years	
Total respondents n=39	1 (2.5%)	11 (28.2%)	11 (28.2%)	5 (13%)	11 (28.2%)	
Sector	Education / HR	Public/ non-profit	Banking/ Finance	Healthcare	Consulting	Technology
Total respondents n=40	7 (17.5%)	3 (7.5%)	5 (12.5%)	1 (2.5%)	4 (10%)	20 (50%)
Area	Europe	UK	Asia	Oceania	North America	Middle East
Total respondents n=40	26 (65%)	8 (20%)	2 (5%)	1 (2.5%)	2 (5%)	1 (2.5%)

Figure B.1: A Summary of the demographics information of surveyed designers (Study 1).

