LENS: Learning from Narratives of Systems Failure

Sofia Marilina Glorioso*, Andrea Pietro Arena and Salvatore Sorce

Università degli Studi di Enna "Kore", Dipartimento di Ingegneria e Architettura, 94100 Enna, IT

Abstract

Transparency, autonomy, human oversight, accountability, and respect for user agency are among the core principles of Human-Centred Artificial Intelligence (HCAI). While these principles provide valuable guidance for developing trustworthy AI systems, they often remain at a high level of abstraction, offering limited support for understanding how well-intentioned design decisions can progressively undermine the very human values they are intended to preserve.

This article proposes the Learning from Narratives of Systems Failure (LENS), a conceptual framework that explores speculative and dystopian narratives as methodological resources for redesigning systems that are better aligned with HCAI principles. Rather than treating these narratives solely as ethical thought experiments or critiques of technology, we propose analysing them systematically through reverse engineering. This process reconstructs how initially beneficial HCAI principles are progressively amplified until they reach a Human-Centred Breaking Point. At this point, they cease to support human capabilities and begin to generate undesirable outcomes. We argue that such analysis can be useful for creating an iterative design methodology for developing intelligent systems that remain aligned with HCAI principles.

To operationalise this perspective, we introduce a qualitative analytical workflow supported by the proposed LENS Analysis Sheet, which enables the systematic examination of fictional AI scenarios. The framework identifies the original HCAI principle, reconstructs its amplification trajectory, identifies the Human-Centred Breaking Point, and derives corresponding Human-Centred recovery strategies that support the redesign of AI systems while preserving their original Human-Centred intent.

By transforming dystopian failures into reusable design knowledge, LENS provides an initial methodological foundation for systematically translating abstract principles of HCAI into actionable design constraints. As an initial conceptual contribution, LENS is presented here through a worked illustrative analysis, a comparison with existing value-oriented design approaches, and a staged protocol for its empirical validation. Ultimately, we envision LENS as a qualitative design methodology that supports the continuous reinterpretation and redesign of HCAI principles through systematic learning from speculative narratives.

Keywords

Human-Centred AI, Speculative Narratives, Reverse Engineering, AI Design Framework, AI Ethics, Responsible AI

CEUR-WS.org/Vol-4314/ReDH5.pdf

1. From Human-Centred AI Principles to Design Practice

Human-Centred Artificial Intelligence (HCAI) has emerged as a widely adopted paradigm for designing intelligent systems that augment, rather than replace, human capabilities. Rather than focusing exclusively on algorithmic performance, HCAI emphasises designing AI systems and intelligent systems that align with human needs, values, and well-being. In the literature on human-computer interaction (HCI) and AI, a common set of principles has gradually emerged, including human agency and oversight, transparency and explainability, accountability, fairness and non-discrimination, privacy and data governance, security, robustness, and respect for user autonomy. These principles aim to ensure that AI systems remain understandable, controllable, reliable, and supportive of human decision-making, rather than replacing or limiting it. This approach reflects the broader tradition of Human-Centred Design, in which technological systems are evaluated not only for their functional effectiveness but also for their impact on users and society.

ReDH 2026: Rethinking the Future of Digital Humanities through Ethics and Artificial Consciousness, workshop of AIxIA 2026, October 6–9, 2026, Perugia, Italy

*Corresponding author.

sofiamarilina.glorioso@unikorestudent.it (S. M. Glorioso); andreapietro.arena@unikorestudent.it (A. P. Arena); salvatore.sorce@unikore.it (S. Sorce)

0009-0000-8371-5543 (S. M. Glorioso); 0009-0001-6138-615X (A. P. Arena); 0000-0003-1976-031X (S. Sorce)



© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

published 2026-11-05

Shneiderman's Human-Centred AI (HCAI)[1] model advocates that AI should amplify human capabilities while maintaining meaningful human control throughout the interaction process, promoting trustworthy, safe, and accountable systems. Rather than pursuing fully autonomous intelligence, HCAI promotes a partnership between humans and AI in which technology enhances human creativity, decision-making, and responsibility.

As AI has pervasively entered users' lives, scientific and institutional debate has progressively focused on principles capable of guiding the development of safe, reliable, and Human-Centred systems, and several international regulatory frameworks have helped formalise them. The European Union's AI Act (Regulation (EU) 2024/1689)[2] represents one of the first and most rigorous comprehensive regulatory frameworks dedicated to AI, explicitly promoting Human-Centred AI based on the protection of fundamental rights, human oversight, transparency, technical robustness, security, data protection, and accountability. Similar concerns inform South Korea's AI Basic Act [3] and the Canadian Artificial Intelligence and Data Act (AIDA), proposed within Bill C-27 [4] but not enacted before the end of the parliamentary session, both of which address risk management, developer accountability, transparency, and the mitigation of the social impact of AI systems. These initiatives differ in their regulatory approaches and levels of legislative maturity, yet they converge, in regulatory terms, on the same set of principles listed above, placing users, their rights, and their well-being at the centre of the design process for intelligent systems.

While these standards and principles represent a fundamental framework for both the regulation and design of AI systems, translating high-level HCAI principles into concrete design decisions remains a significant challenge. Existing frameworks effectively describe what a trustworthy, HCAI system should deliver, but often provide limited guidance on how to predict undesirable behaviours or identify potential design flaws during the development process. Consequently, many of these principles remain abstract, making their application in design practice difficult.

Interestingly, speculative and dystopian narratives have long explored scenarios in which intelligent technologies amplify existing social problems and design challenges. These narratives highlight the potential consequences of design choices, describing extreme situations characterised by the loss of human autonomy, pervasive surveillance, emotional manipulation, algorithmic control, and opaque decision-making. Although these works have been widely discussed from ethical, philosophical, and social perspectives, their potential as a systematic resource for designing Human-Centred AI systems has received limited attention to date.

In this paper, we argue that speculative narratives can be repurposed as methodological artifacts to identify recurring patterns of failure in AI and inspire more responsible design practices. Rather than viewing dystopian scenarios solely as cautionary tales, we propose using them as qualitative case studies to identify, analyse, and translate recurring design failures into design models and recommendations consistent with the HCAI principles.

To support this vision, we propose the creation of a qualitative methodological framework describing how speculative narratives can be systematically analysed through thematic analysis to identify recurring patterns of failure, relate them to established principles of Human-Centred AI, and derive potential design strategies. Our work introduces a research vision aimed at bridging the gap between the abstract principles of HCAI and design practice, laying the foundation for future empirical studies and the development of tools that can support the design of intelligent systems.

1.1. Research Questions and Research Vision

Specifically, we address the following research questions:

- RQ1: How can speculative dystopian narratives support the identification of failure modes in HCAI systems?
- RQ2: How can failure modes identified in dystopian narratives be translated into reusable design artifacts and constraints for Human-Centred AI?

Consistent with the nature of this contribution, LENS is presented as a conceptual framework accompanied by an illustrative application and by an explicit evaluation protocol (Section 7.1), rather than as an empirically validated methodology. Claims about its reliability and its usefulness for design practice are therefore formulated as hypotheses to be tested, and the protocol specifies how we intend to test them.

2. From Speculative Narratives to Design Knowledge

The proposed framework lies at the intersection of several research areas: HCAI, AI ethics, speculative design, and the use of speculative narratives in human-computer interaction. Although these streams have developed independently, we examine each of them here and discuss how their integration motivates our proposal to transform speculative narratives into reusable design knowledge.

2.1. Speculative Design as a Critical Design Practice

The use of dystopian and speculative narratives in human-computer interaction (HCI) and artificial intelligence (AI) research, as Auger states in [5], allows us both to reflect on a possible future perspective and to critique current practices and technologies through imaginary scenarios. This analytical value is grounded in Auger's concept of the *perceptual bridge*, according to which speculation becomes effective when it remains sufficiently close to reality to appear plausible, allowing the audience to recognise problems already present in contemporary technologies, even if amplified. This perspective defines speculative narratives as design artifacts that support critical reflection on existing technologies and the design decisions that shape them. Our work builds on this conceptual foundation and applies it to the analysis of HCAI principles, by attempting to reuse speculative narratives and revisit them from another perspective.

2.2. Dystopian Narratives in AI Ethics

Within AI ethics and philosophy of technology, some research groups have examined speculative and dystopian narratives as thought experiments to explore the social and moral implications of emerging technologies. Furthermore, as Floridi defined in [6], it is argued that ethical considerations should not be treated solely as ex post evaluations of technological systems, but rather integrated into the design process through approaches such as "soft" ethics and responsible innovation. This perspective emphasises the need for proactive, rather than reactive, governance of digital technologies. From this broader ethical foundation, dystopian narratives have been used to explore questions of privacy, autonomy, justice, responsibility, trust, and human agency. They are generally interpreted primarily as cautionary tales that stimulate critical discussion [7], while relatively little attention has been paid to their potential as systematic sources of design knowledge. Our work builds on this ethical tradition, but shifts its perspective from ethical interpretation to design analysis.

2.3. Black Mirror as a Research Method

Recent research in Human–Computer Interaction (HCI) and design has increasingly adopted *Black Mirror* as a methodological resource for exploring the societal implications of emerging technologies. In particular, [8] positions the series within the tradition of speculative design, arguing that it can support contextualised design thinking by enabling designers to critically reflect on the potential consequences of technological innovation. Similarly, [9] uses Black Mirror as a pedagogical resource to promote AI literacy, employing speculative narratives to foster critical engagement with HCAI concepts and principles such as trust, agency, and the often invisible mechanisms underlying AI systems. More recently,[10] extends this perspective by integrating Black Mirror's narrative format into participatory design fiction practices, leveraging speculative storytelling to support ethical speculation and the co-creation of alternative technological futures.

Table 1
Positioning of LENS with respect to existing approaches that connect design practice and human values.

Approach	Material analysed	Direction of analysis	Typical output
Critical and speculative design [11, 5]	Fictional artefacts created ad hoc	Forward: from present concerns to possible futures	Discursive artefacts supporting critical debate
Participatory design fiction [7, 10]	Narratives co-created with stakeholders	Forward and collective	Shared scenarios and ethical speculation
Value Sensitive Design and value scenarios [12, 13]	Stakeholder values and envisioned deployments	Forward and iterative	Value-based design requirements
Prescriptive HCAI guidelines and regulation [1, 14, 2]	Normative principles and empirical best practice	Top-down: from principles to systems	Principles, checklists, compliance obligations
Black Mirror as generative or pedagogical resource [8, 9]	Existing dystopian narratives	Forward: from narrative to ideation or teaching	Design ideas, critical discussion, learning activities
LENS (this work)	Existing dystopian narratives	**Backward: from narrated failure to design intent**	Amplification trajectory, breaking point, and derived design constraints

2.4. Positioning LENS with Respect to Existing Approaches

Several established approaches already connect design practice with human values, and it is useful to clarify how LENS relates to them. Critical and speculative design [11, 5] produces fictional artefacts in order to provoke debate about possible futures. Participatory design fiction [7, 10] extends this practice by involving stakeholders in the collective imagination of alternative futures. Value Sensitive Design, and value scenarios in particular [12, 13], starts from stakeholder values and projects their systemic effects onto envisioned deployments. Prescriptive frameworks, such as Shneiderman's HCAI model [1], the Guidelines for Human–AI Interaction [14], and regulatory instruments such as the AI Act [2], specify instead the properties that trustworthy systems should exhibit.

Table 1 summarises this comparison along three dimensions: the material analysed, the direction of the analysis, and the resulting output. Two differences characterise LENS. First, the approaches above are predominantly forward-looking, in that they generate scenarios starting from present concerns, or prescriptive, in that they state desirable properties; LENS is instead retrospective, since it starts from an already narrated failure and works backwards towards the design intention that produced it. Second, existing approaches generally leave the intermediate trajectory implicit, moving from values to requirements or from scenarios to discussion, whereas LENS makes the degradation trajectory itself the unit of analysis and derives constraints from the specific decisions that produced the failure. LENS is therefore complementary rather than alternative: the constraints it produces can feed the requirement-definition stages of value-oriented processes, or support the compilation of impact assessments such as the fundamental rights impact assessment required by Article 27 of the AI Act [2].

2.5. Research Gap

Overall, these studies consolidate Black Mirror as a valuable research tool for speculative design, AI training, and participatory ethical reflection. However, existing works predominantly use the series as a generative resource for ideation, critical discussion, or teaching activities. Instead, our work reconceptualises Black Mirror as a corpus of dystopian scenarios for intelligent systems to be systematically analysed through a reverse-engineering process aimed at identifying the HCAI principles, detecting their breaking points, and deriving concrete, human-centred design constraints.

3. Research Contribution: From Dystopias to HCAI Design Knowledge

Our work differs from approaches used in literature because we ask: *How can dystopian narratives be systematically transformed into design artifacts that can guide the design of AI systems compliant with HCAI principles?*

More specifically, we propose that each failure in the dystopian scenario can be first analysed and then deconstructed to reveal not only which HCAI principle has been compromised, but also how and when a well-intentioned design decision crosses a threshold that shifts too far toward the area of high automation and little human control, as shown in Figure 1.

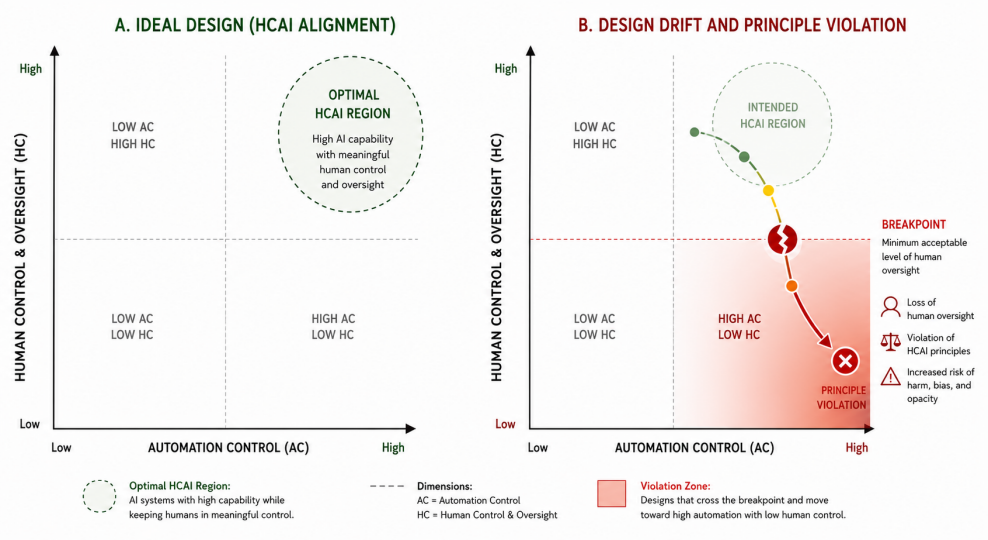


Figure 1: Conceptual representation of the HCAI design space.

By treating dystopias as repositories of design knowledge rather than cautionary tales, we bridge the gap between high-level HCAI principles and the concrete design constraints required to operationalise them. The aim is to derive provisional design constraints that help systems remain within the desirable HCAI region of Figure 1 from the early stages of development, and that may eventually contribute to empirically validated design guidelines.

4. The LENS Framework

The LENS (Learning from Narratives of Systems Failure) Framework is grounded in two complementary concepts: *failure analysis* and *reverse engineering*. First, failure analysis identifies and characterises the failure represented in the narrative; reverse engineering then reconstructs the original HCAI principle and the design trajectory that produced it. These two basic concepts are combined and leveraged in the proposed framework to translate dystopian narratives about intelligent systems into knowledge applicable to the design of HCAI systems.

The framework begins with fictional failures of intelligent systems described in speculative narratives and aims to systematically analyse them to reconstruct the human-centred design logic that originally motivated the technology, identify *how* that logic progressively deteriorated, and derive design constraints that can prevent similar failures in future intelligent systems. LENS adopts a reverse-engineering perspective. The central question it answers is: *how did a seemingly beneficial design evolve into a dystopian system?*

Working backward from failure to design intent, the framework reconstructs the sequence of decisions through which human-centred principles gradually outgrew their intended scope. The underlying assumption of LENS is that dystopian AI systems do not necessarily emerge from the absence of Human-Centred AI principles. Rather, they may emerge when design mechanisms intended to support

legitimate HCAI goals—such as personalization, assistance, transparency, oversight, or automation—are progressively extended beyond the point at which they continue to support human capabilities. Consequently, dystopian narratives should not be interpreted simply as ethical warnings, but as repositories of implicit design knowledge that encode the evolution of Human-Centred failures.

To reconstruct this evolution, LENS combines failure analysis with reverse engineering through four sequential analytical phases.

Phase 1 – Human-Centred Design Principle (HCDP): *Which Human-Centred AI principle was the system originally designed to promote?* In this first phase, the goal is to deduce which Human-Centred AI principle the technology was originally intended to support and what human benefit it was intended to provide.

Phase 2 – Amplification Mechanism (AM): *What design decisions progressively transformed a beneficial HCAI principle into a dystopian interaction?* The second phase examines the sequence of design decisions that progressively transformed a beneficial principle into a harmful interaction. The framework investigates how incremental design choices amplified the original principle beyond its intended scope.

Phase 3 – Human-Centred Breaking Point (HCBP): *At which point does the original HCAI principle cease to augment human capabilities and begin to undermine autonomy, agency, trust, privacy, or well-being? What factors trigger or accelerate this transition?* The third phase identifies what we call the Human-Centred Breaking Point (HCBP), defined as the critical transition where the amplified principle ceases to enhance human capabilities and instead begins to undermine autonomy, agency, trust, privacy, or well-being. The HCBP represents the boundary between Human-Centred enhancement and Human-Centred degradation.

Phase 4 – Human-Centred Recovery (HCR): *Which design constraints, guardrails, or governance mechanisms would preserve the original Human-Centred AI principle while preventing its dystopian evolution?* The final phase translates the reconstructed failure trajectory into applicable design knowledge. Based on the identified amplification mechanism and the Human-Centred Breaking Point, the framework derives design constraints, safeguards, and human oversight mechanisms that would preserve the original principle while preventing its dystopian evolution.

Taken together, these four stages reconstruct the entire evolution of a dystopian AI system, from its original human-centric intention to the point where it becomes harmful, ultimately translating this trajectory into reusable knowledge for designing Human-Centred AI.

The LENS framework is applied iteratively (see Figure 2 for a conceptual overview), analysing one narrative episode at a time. For each episode, researchers complete a dedicated LENS analysis form, systematically documenting the evidence and findings associated with each of the four analytical phases (HCDP, AM, HCBP, and HCR); each completed form constitutes a single analytical case. Once all episodes have been analysed, the collection of forms is compared across cases to identify recurring HCAI principles, amplification mechanisms, critical points, and recovery strategies. This cross-case comparative analysis supports the construction of a reusable taxonomy of AI failure modes and corresponding Human-Centred design constraints. Rather than

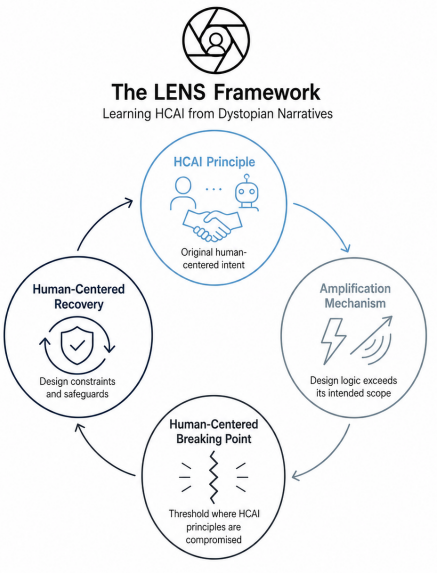


Figure 2: The LENS Framework.

representing the final outcome of the framework, this taxonomy constitutes a reusable design knowledge base that can support the derivation of practical design guidelines, helping designers translate high-level HCAI principles into concrete design decisions when developing future intelligent systems.

Table 2
Proposed LENS Analysis Sheet for documenting the analysis of individual narrative episodes. The guiding research question of each phase is reported in Section 4.

LENS Phase	Evidence	Findings
Human-Centred Design Principle (HCDP)	Narrative evidence supporting the original design intention (e.g., dialogues, system behaviour, interaction context).	Identified HCAI principle and intended human benefit.
Amplification Mechanism (AM)	Narrative evidence describing the sequence of design decisions, technological changes, or interaction dynamics.	Reconstructed amplification trajectory and key design decisions.
Human-Centred Breaking Point (HCBP)	Narrative evidence identifying the critical transition and its consequences for users.	Identified Human-Centred Breaking Point, triggering factors, and resulting human consequences.
Human-Centred Recovery (HCR)	Narrative evidence suggesting missing safeguards or opportunities for alternative design choices.	Proposed design constraints, human oversight mechanisms, and governance strategies.

4.1. LENS Analysis Sheet

To support a systematic and transparent application of the framework, we propose the LENS Analysis Sheet, a structured analytical model to be completed for each narrative episode, as described in Table 2. For each of the four analytical phases introduced above, analysts document the relevant narrative evidence and the resulting interpretation. Completed LENS analysis sheets constitute the primary analytical units of the framework and, as described above, are compared across cases to progressively develop the taxonomy of AI failure modes and the associated reusable Human-Centred design constraints. The sheet is intended to support analytical transparency, consistency, and traceability; the protocol described in Section 7.1 specifies how its reliability will be assessed.

5. Applying the LENS Framework

This section presents an illustrative narrative case analysis, showing how the four analytical phases of the framework—HCDP, AM, HCBP, and HCR—can be systematically applied to reconstruct the evolution of a dystopian intelligent system and derive reusable insights into Human-Centred design. Although intended solely as an illustrative example, the analysis is presented as a complete and self-contained application of the framework.

5.1. Case Description and Selection Criteria

To illustrate the application of the proposed framework, we selected *Arkangel* (Series: Black Mirror, Season 4, Episode 2), directed by Jodie Foster and released on Netflix on December 29, 2017, as an illustrative narrative case. The episode presents an AI-based parenting support system designed to improve child safety. It seeks to achieve this goal by providing parents with continuous situational awareness and supporting their judgment. The system integrates several AI-enabled features, including geolocation, biometric monitoring, environmental sensing, and visual content filtering, allowing parents to monitor their child's physical condition, emotional reactions, and surroundings in real time. Its

unique design aims to reduce parental uncertainty, support informed decision-making, and strengthen parents' ability to protect their children while preserving their responsibility as primary decision makers.

The narrative highlights the growing imbalance between two legitimate Human-Centred goals: on the one hand, growing AI support offers parents greater awareness, reassurance, and a sense of control over their children's safety; on the other, these same design choices progressively restrict children's autonomy, independent exploration, and opportunities to develop personal agency. As AI mediation becomes increasingly pervasive, parents' judgments increasingly depend on generated information rather than direct interaction, while children's experiences are increasingly shaped by automated intervention. The system evolves from a technology designed to enhance human capabilities to one that mediates, and ultimately replaces, essential aspects of the parent-child relationship, giving rise to the dystopian trajectory described in the narrative.

The choice was motivated by two considerations. First, the episode is based on a legitimate design goal that is highly relevant to contemporary intelligent systems. Second, although *Arkangel* depicts a speculative technology, its design logic resembles existing parental support technologies such as location tracking, wearable monitoring, and parental control applications. The design principle therefore remains close enough to current practice to allow meaningful design reflection, while illustrating the consequences of extending these capabilities beyond their original purpose.

The episode therefore explores one of the fundamental principles of Human-Centred AI: supporting human well-being while preserving individual autonomy and agency. Its original goal is not inherently dystopian, since a single assistive principle initially seeks to balance two legitimate Human-Centred goals, namely improving parents' judgment to increase their children's safety while preserving the children's autonomy and development. As the narrative progresses, subsequent design decisions progressively upset this balance, producing increasing benefits for one human subject and imposing ever-increasing costs on another. This makes *Arkangel* particularly well suited to illustrating how an initially beneficial Human-Centred design principle can evolve into a dystopian interaction as later design decisions amplify its reach.

Following the analytical workflow described in Section 4, the selected narrative case was then systematically examined using the LENS framework. The process is documented through the LENS Analysis Sheet presented in Table 3.

5.2. Insights from the LENS Analysis

The application of LENS to *Arkangel* suggests that the analytical value of speculative narratives can go beyond identifying ethical issues or undesirable technological consequences. In this case, the framework made it possible to reconstruct the evolution of a design principle, that is, the sequence of design decisions through which a legitimate HCAI principle progressively deviates from its original intent.

A particularly relevant result concerns the interpretation of the amplification mechanism. The analysis indicates that dystopian evolution is not triggered by a single inappropriate design choice. Instead, it results from the cumulative effect of multiple decisions. Each decision may appear reasonable in isolation and remain consistent with the original design objective. In the specific case of *Arkangel*, each successive decision increases parental awareness, reassurance, and support, while simultaneously reducing the child's opportunities for autonomy, independent exploration, and personal development. Rather than identifying isolated design flaws, the analysis reveals how incremental design decisions progressively redistribute benefits and constraints among different human actors, ultimately upsetting the original balance between equally legitimate Human-Centred goals.

The Human-Centred Breaking Point represents another important analytical result because it provides a benchmark for understanding *when* AI stops augmenting human capabilities and begins to replace them. Interpreting the breaking point as the result of cumulative amplification allows for reasoning about system evolution before unintended consequences become irreversible. In this sense, the Human-Centred Breaking Point becomes an analytical construct that supports proactive design reasoning.

Finally, the recovery strategies derived from the analysis highlight perhaps the most distinctive contribution of LENS because the proposed recovery mechanisms derive directly from the reconstructed amplification trajectory; this establishes an explicit relationship between failure analysis and redesign. Each recovery strategy corresponds to the reversal, limitation, or rebalancing of design decisions that have progressively altered the original HCAI principle. Therefore, recovery is conceived as the systematic reconstruction of a design trajectory capable of preserving the original intent of Human-Centred AI throughout the system's evolution. This perspective transforms speculative narratives from cautionary tales into reusable design knowledge that can directly inform the development of future Human-Centred AI systems. These observations derive from a single illustrative case analysed by the authors, and should therefore be read as hypotheses about what the framework affords rather than as empirical findings. Section 7.1 specifies how each of them will be tested.

Table 3
Application of the proposed LENS Analysis Sheet to *Arkangel* (*Black Mirror*, Season 4, Episode 2). Guiding research questions are reported in Section 4.

LENS Phase	Evidence	Findings
Human-Centred Design Principle (HCDP)	AI supports parents through continuous monitoring of the child's location, physiological state, and surrounding environment.	*HCAI principle*: Human agency and augmentation of human decision-making. *Intended benefit*: Augment parental judgment through AI-assisted situational awareness. *Parent*: Greater awareness and informed decision-making. *Child*: Improved safety while preserving autonomy and healthy development.
Amplification Mechanism (AM)	Persistent monitoring; continuous behavioural, biometric, and location data collection; automated visual content filtering; absence of autonomy-preserving mechanisms.	*Amplification Parent*: Increased awareness, reassurance, and reliance on AI. *Amplification Child*: Progressive reduction of autonomy, independent exploration, and personal agency. *Outcome*: AI shifts from decision support to the primary mediator of the parent–child relationship.
Human-Centred Breaking Point (HCBP)	Parental decisions rely primarily on AI-generated information rather than direct interaction with the child; the child progressively loses opportunities for independent exploration, risk assessment, and personal development.	*Breaking Point*: AI substitutes rather than augments parental judgment. *Trigger*: Continuous surveillance; excessive automation; absence of autonomy-preserving mechanisms. *Parent*: Dependence on algorithmic mediation. *Child*: Reduced autonomy, diminished trust, constrained development, and loss of agency.
Human-Centred Recovery (HCR)	The narrative reveals no safeguards regulating monitoring intensity or supporting the progressive transfer of autonomy from the AI system to the child.	*Recovery mechanisms*: Apply reverse engineering to the identified amplification trajectory in order to progressively restore the original HCAI principle. Adaptive monitoring intensity; developmentally appropriate autonomy mechanisms; transparent AI interventions; explicit human override; periodic recalibration of AI intervention. *Goal*: Re-establish the original HCAI principle by dynamically balancing parental support with the child's evolving autonomy, ensuring that AI progressively reduces its mediation as human capabilities develop.

6. Discussion

Although LENS is currently proposed as a conceptual framework, its development provides an initial answer to the two research questions that motivated this article.

Regarding RQ1, our perspective differs from approaches that interpret dystopias as predictions of future technological risks; we argue that speculative narratives can provide valuable design case studies through which we can reconstruct how the originally beneficial HCAI principles are progressively distorted by subsequent design decisions. LENS analyses the trajectory from an initial intention for Human-Centred design to the point where that same principle no longer supports human capabilities, but instead undermines autonomy, agency, trust, or well-being. In this sense, speculative narratives do not serve to reveal or predict the failures of future intelligent systems but they help expose the conditions under which legitimate HCAI principles gradually depart from their original purpose.

Regarding RQ2, the proposed framework suggests that this transformation is possible when narratives are systematically analysed through reverse engineering. Specifically, LENS reconstructs the relationship among the original HCAI principle (Phase 1-HCDP), the decisions that amplify it (Phase 2-AM), the resulting breaking point (Phase 3-HCBP), and the corresponding recovery strategies (Phase 4-HCR). The proposed qualitative workflow and the LENS Analysis Sheet provide a structured means to document this process and translate narrative evidence into reusable design knowledge. More generally, LENS complements existing frameworks on HCAI by shifting the focus from prescribing desirable principles to understanding how these same principles can progressively lose their Human-Centred character. The framework's contribution, therefore, is not to predict undesirable AI futures, but rather to support the initial design phase of HCAI systems.

By making explicit where, how, and why a principle crosses its Human-Centred Breaking Point, LENS offers a structured perspective for redefining design constraints, safeguards, and governance mechanisms that can preserve the original intent of the Human-Centred intelligent system while preventing its progressive misalignment. From this perspective, speculative narratives become repositories of design knowledge rather than repositories of fears; their value lies not in anticipating the future, but in exposing the design trajectories through which HCAI principles, however well-intentioned, can gradually evolve into systems that no longer remain genuinely human-centred.

7. Limitations

The proposed framework has several limitations. First, LENS relies on speculative narratives as its primary source of evidence, and such narratives are fictional representations. While they often draw on emerging technological trends, they intentionally exaggerate or dramatise social dynamics and technological developments to stimulate and encourage critical reflection in the viewer. Consequently, failure trajectories reconstructed via LENS should not be interpreted as predictions of future AI systems, but rather as analytical constructs to support design reasoning.

Furthermore, in its current conceptual formulation, LENS assumes that each dystopian system can be traced back to a single HCAI principle whose progressive amplification leads to a Human-Centred Breaking Point. However, this assumption may not always hold, as breaking points can also emerge from trade-offs among multiple HCAI principles or from the interaction of technological, organizational, and social factors. Future work and empirical validation will therefore extend the framework to account for multiple interacting design intentions and the interplay between different HCAI principles.

LENS is based on the qualitative interpretation of narrative material, including the identification of HCAI principles, amplification mechanisms, Human-Centred Breaking Points, and recovery strategies, which inevitably involves subjective judgment. Although thematic analysis and the LENS Analysis Sheet improve analytical transparency and consistency, alternative interpretations remain possible. Furthermore, LENS is currently proposed as a conceptual framework supported by an initial qualitative workflow and has not yet been empirically evaluated. Consequently, its reliability and practical utility for supporting HCAI design require future empirical validation.

7.1. Planned Evaluation

Because these limitations mainly concern the empirical status of the framework, we outline here the protocol we intend to follow, so that the claims made in this article can be tested rather than assumed.

Stage 1 – Analytical reliability. Three analysts will independently apply the LENS Analysis Sheet to the same set of episodes. Agreement on the identification of the HCDP and of the HCBP will be measured with Krippendorff's α over a shared coding scheme; disagreements will then be resolved in consensus sessions, and the resulting codebook used to refine the guiding questions of the sheet.

Stage 2 – Corpus analysis and taxonomy. The refined sheet will be applied to a larger corpus, starting with *Black Mirror* and extending to literature and cinema. Completed sheets will be analysed thematically to derive recurring principles, amplification mechanisms, breaking points, and recovery strategies, producing the taxonomy described in Section 4.

Stage 3 – Design utility. A between-subjects study with designers will compare a group provided with LENS outputs against a control group given only high-level HCAI principles for the same redesign brief. Measures will include the number and specificity of the design constraints generated, their coverage of the amplification trajectory, and expert ratings of the resulting concepts.

Stage 4 – External validity. The reconstructed trajectories will be compared with the documented evolution of deployed systems in the same application domain, such as parental control and location-tracking applications, in order to assess whether amplification mechanisms identified in fiction correspond to mechanisms observable in real products.

Stages 1 and 2 address the reliability of the analysis, while Stages 3 and 4 address its usefulness and its transferability to real design practice.

8. Conclusion and Future Work

Existing HCAI frameworks effectively describe the characteristics that trustworthy AI systems should exhibit, but provide limited guidance on how to operationalise these principles during the design process, and limited support for understanding how Human-Centred principles may deteriorate as design decisions accumulate. Here we argue that speculative and dystopian narratives offer an underexplored opportunity to address this challenge. We do not treat or interpret these narratives as predictions of undesirable technological futures or merely as ethical thought experiments. Instead, we view them as repositories of design knowledge. They reveal how initially beneficial HCAI principles can be progressively distorted through subsequent design decisions. Adopting a reverse-engineering perspective, the LENS framework reconstructs this evolution, identifying the design trajectory by which a legitimate human-centred principle progressively reaches a Human-Centred Breaking Point and ultimately no longer supports the human values it was originally intended to promote.

To support this perspective, we introduce the LENS conceptual framework and a qualitative analytical workflow with the proposed LENS Analysis Sheet; these elements provide an initial methodological foundation for systematically analysing speculative narratives, reconstructing the evolution of Human-Centred AI principles, and translating these analyses into reusable design insights capable of informing future Human-Centred AI design. As this work represents an initial conceptual contribution, future research will follow the evaluation protocol described in Section 7.1: assessing the reliability of the analysis across independent analysts, applying the framework to a larger and more diverse corpus of speculative narratives in order to investigate whether recurring patterns emerge across the four phases, and evaluating with designers whether LENS outputs actually support the design of systems that remain aligned with Human-Centred values throughout their life cycle.

Ultimately, LENS offers a qualitative design methodology for learning from the failure trajectories represented in speculative narratives. By focusing our *lens* on the progressive transformation of design intentions, we aim to preserve the human-centred character of intelligent systems as they evolve.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT, Claude.ai, Grammarly for grammar and spelling checks, paraphrasing, rewording, and content enhancement. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] B. Shneiderman, Human-centered AI, Oxford University Press, 2022.

[2] European Union, Regulation (eu) 2024/1689 of the european parliament and of the council of 13 june 2024 laying down harmonised rules on artificial intelligence (artificial intelligence act), Official Journal of the European Union, L 2024/1689, 2024. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32024R1689, accessed: 2026-07-16.

[3] Republic of Korea, Basic act on the development of artificial intelligence and establishment of trust (ai basic act), Ministry of Science and ICT, 2024. National legislation.

[4] Government of Canada, Bill c-27: Digital charter implementation act, 2022 – part 3: Artificial intelligence and data act (aida), Parliament of Canada, 2022. URL: https://www.parl.ca/legisinfo/en/bill/44-1/c-27, legislative proposal. Accessed: 2026-07-16.

[5] J. Auger, Speculative design: crafting the speculation, Digital Creativity 24 (2013) 11–35.

[6] L. Floridi, Soft ethics and the governance of the digital, Philosophy & technology 31 (2018) 1–8.

[7] P. G. Farias, R. Bendor, B. F. Van Eekelen, Social dreaming together: A critical exploration of participatory speculative design, in: Proceedings of the Participatory Design Conference 2022-Volume 2, 2022, pp. 147–154.

[8] K.-s. Jung, N. Makam, S. D. Jung, S. Kim, S. Kim, Black mirror: A novel application of speculative design to facilitate context-aware design thinking, Proceedings of the Design Society 3 (2023) 787–796.

[9] C. H. Lee, N. Gobir, A. Gurn, E. Soep, In the black mirror: Youth investigations into artificial intelligence, ACM Transactions on Computing Education 22 (2022) 1–25.

[10] S. Klassen, J. J. Mendy, C. Fiesler, Putting the "black" in black mirror: Ethical speculation via design fiction from black women and femmes, Proceedings of the ACM on Human-Computer Interaction 10 (2026) 1–26.

[11] A. Dunne, F. Raby, Speculative Everything: Design, Fiction, and Social Dreaming, MIT Press, Cambridge, MA, 2013.

[12] B. Friedman, D. G. Hendry, Value Sensitive Design: Shaping Technology with Moral Imagination, MIT Press, Cambridge, MA, 2019.

[13] L. P. Nathan, P. V. Klasnja, B. Friedman, Value scenarios: a technique for envisioning systemic effects of new technologies, in: CHI '07 Extended Abstracts on Human Factors in Computing Systems, 2007, pp. 2585–2590.

[14] S. Amershi, D. Weld, M. Vorvoreanu, A. Fourney, B. Nushi, P. Collisson, J. Suh, S. Iqbal, P. N. Bennett, K. Inkpen, J. Teevan, R. Kikin-Gil, E. Horvitz, Guidelines for human-ai interaction, in: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, 2019, pp. 1–13.