

Aligning With the Dead: Posthumous Persistence as a Limit Case for Bidirectional Human-AI Alignment

Ruben Schlonsak

Technical University of Applied Sciences Lübeck
Fraunhofer IMTE, Lübeck, Germany
ruben.schlonsak@th-luebeck.de

Denys Matthies

Technical University of Applied Sciences Lübeck
Fraunhofer IMTE, Lübeck, Germany
denys.matthies@th-luebeck.de

Abstract

Bidirectional alignment assumes that both parties, human and AI, continue to evolve. But what happens when one side stops? This position paper treats posthumous persistence as a stress test for alignment research. Building on the Bidirectional Human-AI Alignment framework [11] and concrete contexts, longitudinal health monitoring, personal AI assistants, and digital legacy systems, we argue that a user’s death breaks several quiet premises in current alignment accounts, that values remain negotiable, that consent can be renewed, and that human in the loop implies a living human. We introduce three design scenarios to make these fractures explicit. From them, we derive three mechanisms aimed at the temporal boundary of bidirectional alignment: Alignment Advance Directives, Alignment Sunset Protocols, and Inheritor Onboarding Interfaces. Taken together, these mechanisms shift alignment from an indefinitely ongoing negotiation to something that can be specified, bounded, and transferred. Our broader claim is methodological: HCI should not treat posthumous alignment as a niche concern. It is a diagnostic lens, one that reveals whether a bidirectional alignment approach is robust when continuity, consent, and agency are no longer guaranteed.

CCS Concepts

• **Human-centered computing** → **HCI theory, concepts and models.**

Keywords

posthumous data, bidirectional alignment, value persistence, digital legacy, speculative design, health sensing

1 Introduction

Human-AI alignment is increasingly understood as more than a one-time calibration. It is a dynamic, bidirectional process [4, 11]. Humans shape AI through feedback, stated preferences, and interaction habits. AI, in turn, shapes humans through recommendations, nudges, and the incremental restructuring of what options feel available or likely [2]. This interchange assumes two active partners.

We propose a simple thought experiment: *What happens to this process when one partner dies?*

Contemporary AI systems accumulate months or years of interaction data, preference inferences, and behavioral traces [9]. Health monitoring platforms collect continuous biometric streams that persist long after their wearers; personal AI assistants build increasingly specific models of individual users over time. When that user dies, these systems enter a condition they were not designed

for: alignment without a living counterpart. We argue that posthumous persistence, the continued existence of AI systems, data, and user models after the human counterpart’s death, is a limit case for bidirectional alignment. It exposes assumptions usually left implicit: that values remain revisable, consent can be renewed, and the human in the loop is alive. By focusing on what fails at this temporal boundary, we identify where alignment theories are brittle, and move from diagnosis to design by sketching three concrete mechanisms to make alignment more resilient to human mortality.

2 Related Work

2.1 Death and Technology in HCI

HCI research on death and technology has examined digital memorialization [7], how social media profiles persist after death [1], and the rise of the digital afterlife industry [8]. Legal scholarship has also addressed how digital assets can be transferred or inherited [5]. Much of this work, however, treats the deceased user’s data as a *static artifact*, something to preserve, delete, or hand over. It rarely confronts a key shift in contemporary systems: many AI-enabled products are not passive archives. They continue to learn, infer, and act on the basis of accumulated interaction histories.

Related theoretical traditions also tend to assume continued human participation. Value-sensitive design [3] offers methods for embedding human values in technology, but it presumes that stakeholders remain available to negotiate and revise those values over time. Activity theory [6] emphasizes that human-technology interaction is situated and evolving, yet those dynamics end abruptly at death. Within AI alignment research [4], there is growing recognition that value specifications must be pluralistic and changeable [12]. Even so, the temporal limits of value validity, and what happens when values can no longer be updated by the person they came from, remain underexplored.

2.2 Bidirectional Alignment and Its Temporal Blind Spot

Shen et al. [11] propose the Bidirectional Human-AI Alignment framework, which frames alignment as two coupled directions, “Aligning AI with Humans” and “Aligning Humans with AI.” Based on a systematic review of more than 400 papers across HCI, NLP, and ML, they articulate four core research questions that span human value specification, integrating values into AI systems, human cognitive adjustment to AI, and human adaptive behavior. The framework describes alignment as a continuous feedback process in which both sides mutually adapt.

Like much of the literature it synthesizes, this framing assumes that both sides of the loop remain available as active participants. In

particular, Shen et al.'s discussion of value specification techniques (RQ1) centers on three channels: explicit human feedback, implicit behavioral signals, and simulated human value feedback [11]. Each ultimately depends on a living source of values, behavior, or ground truth. Their account of co-evolution further emphasizes the need for ongoing oversight and updates, including anticipating how human values may evolve over time. The framework does not ask what it means for that evolution to stop.

Several gaps identified by Shen et al. point directly to the boundary case we study. They note that implicit and simulated value feedback remain relatively underexamined, and we argue that this is partly due to structural factors. Both channels presuppose temporal continuity that death breaks. They also highlight that deployment-time personalization for individuals is underexplored, and our scenarios show why that matters. Personalization can produce relational artifacts that are meaningful precisely because they are tied to a particular person and are difficult to transfer. Finally, they emphasize the lack of research on long-term interaction. We extend that concern by asking what "long-term" entails when one participant's timeline ends.

Our contribution is to make this temporal blind spot explicit. We extend the bidirectional alignment framing by examining what happens when the human side of the loop permanently ceases to provide feedback, revise values, or adapt behavior.

3 What Breaks When the Human Dies

Current approaches to bidirectional alignment implicitly rely on at least three conditions, each of which death breaks. Using Shen et al.'s framework as an organizing lens, we map these conditions onto specific components of their model [11]:

Continuous consent. In Shen et al.'s framework, human value specification (RQ1) is sustained by ongoing interaction, via explicit feedback (ratings, instructions, natural language), implicit feedback (behavioral cues, physiological signals), and simulated feedback intended to approximate human responses [11]. Each channel presupposes a living, responsive person who can revisit, clarify, or revise prior choices. Death ends that loop. Preferences become fixed at their last observable state, independent of how they might have shifted later. This is not a minor technical gap, it contradicts the underlying value model itself, since Schwartz's Theory of Basic Values characterizes values as priorities that change across life stages [10], a trajectory that death terminates.

Mutual adaptation. Bidirectional alignment hinges on reciprocal change, systems adapt to humans (RQ1–RQ2), and humans adapt to systems (RQ3–RQ4). After death, that reciprocity collapses. The system can continue to update, through new training data, policy revisions, deployment changes, or platform-level updates, while the human side of the relationship cannot respond at all. This creates a structural asymmetry where the alignment target is static, but the optimizer is not. In this sense, posthumous persistence is an extreme case of the misalignment dynamics observed when reward functions shift over time [2], except that the "human reward function" is no longer merely drifting, it is absent. The system can only move away from its last anchoring signal.

Identifiable stakeholders. Death also destabilizes the question of who the system is for. Once the user dies, the number of

candidate beneficiaries grows, including the deceased's last-known wishes, family members, clinicians, researchers, and the platform itself, often with incompatible interests. Shen et al. explicitly recognize plural stakeholders and motivate pluralistic value alignment via social choice theory [12]. However, their framing still assumes that the primary stakeholder can participate in resolving value conflicts. Posthumous alignment turns stakeholder resolution into a governance problem with an absent principal, meaning conflicts must be adjudicated without direct participation from the person whose values originally grounded the system. Figure 1 summarizes how death transforms the bidirectional alignment loop into a unidirectional condition in which all three assumptions break.

4 Three Design Scenarios

To make these tensions tangible, we outline three near-future scenarios. Each illustrates a distinct posthumous alignment failure mode, and each maps to a gap in Shen et al.'s bidirectional alignment framework.

4.1 The Silent Insole

Maria, 67, is a colorectal cancer survivor enrolled in a longitudinal health monitoring study. She wears a smart insole that records continuous gait, activity patterns, and physiological signals using embedded IMU and plantar pressure sensors. Over 14 months, an AI health companion learns her routines, flags subtle gait changes, and offers personalized rehabilitation nudges. Maria then dies unexpectedly from an unrelated cardiac event.

The system still holds 14 months of high-frequency biometric data. The research platform has also derived a detailed model of her movement patterns, potentially useful for detecting early decline markers in other patients. Maria consented to research use, but her consent was framed around *her own* care and benefit, not about her data being used to train models that would shape the behavior of systems deployed to strangers. After her death, the alignment relationship that governed how the system should treat "Maria's data" can no longer be renegotiated.

Alignment tension: The system was aligned to serve Maria, through a live feedback relationship. Posthumously using her data to help others may be socially beneficial, yet it shifts the normative target: the system is no longer optimizing for Maria's lived interests, but for downstream populations and institutional goals. In Shen et al.'s terms, the implicit feedback channel (behavioral and physiological signals) has been irreversibly severed, while the artifacts produced by that channel continue to circulate. Who, then, is authorized to define "aligned" behavior toward Maria's data: Maria's last-known values, her family's preferences, or a broader public-interest calculus?

4.2 The Evolving Assistant

Jens, a 42-year-old researcher, has used an LLM-based writing assistant every day for three years. Through what Shen et al. call *interactive alignment*, using steady back and forth feedback, steering prompts, and personalization driven by his usage history, the system has come to mirror his writing voice, anticipate the themes he returns to, and align its suggestions with the professional outcomes he consistently works toward[11]. Jens dies in an accident.

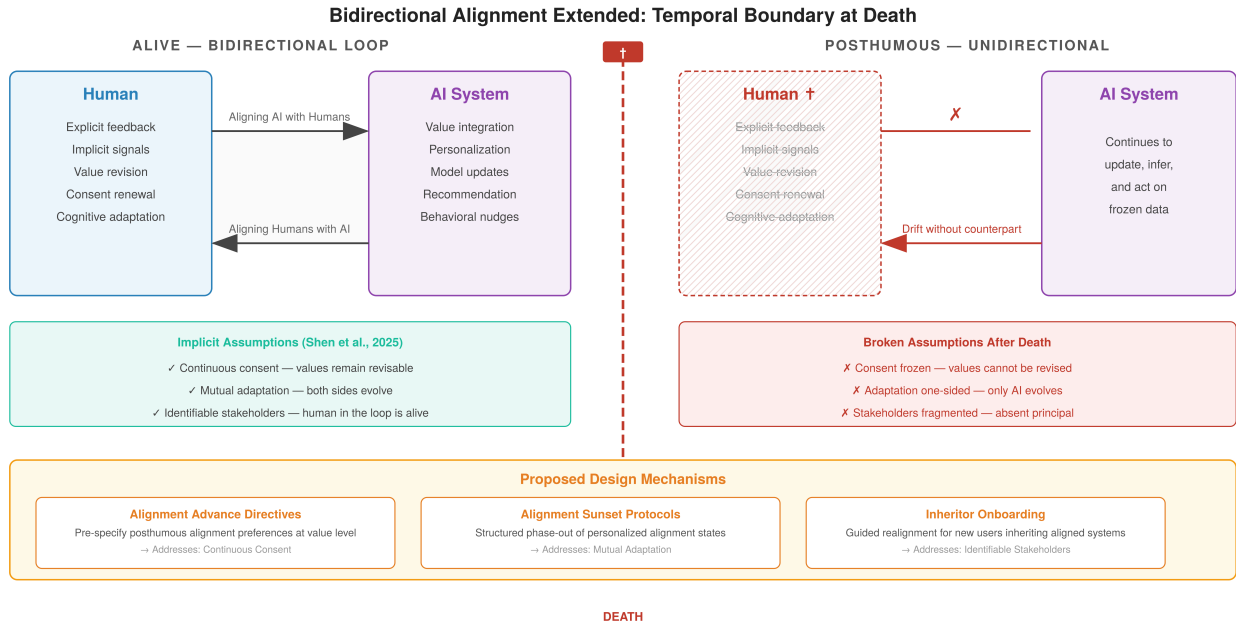


Figure 1: Bidirectional alignment extended to the temporal boundary of death. Left: the living feedback loop between human and AI, sustained by continuous consent, mutual adaptation, and identifiable stakeholders. Right: after death, the human side permanently exits the loop, breaking all three assumptions while the AI system continues to evolve.

Six months later, the platform deploys a major model update that substantially changes how the assistant interprets prompts and applies learned preferences.

Jens’s university wants access to his interaction logs to reconstruct an unfinished grant proposal. But the updated model now reads Jens’s old prompts through a different decision procedure than the model Jens actually worked with. The resulting “completed” document reflects a hybrid: Jens’s partial intentions, filtered through a system configuration he never used and never consented to.

Alignment tension: Bidirectional alignment here is not just person-specific, it is *version-specific*. It depended on Jens interacting with a particular model at a particular time. After death, only one side can change. The assistant continues to evolve; Jens cannot. Shen et al. emphasize “continuously updating AI with limited data without compromising alignment values” [11], but this case has *no* new data at all, only a frozen interlocutor and a moving model. The scenario exposes a temporal signature of alignment that current frameworks do not preserve, suggesting that alignment may require provenance not only of data but also of model versions and policy contexts.

4.3 The Inherited Dashboard

Ayumi, 78, managed diabetes using an AI health dashboard integrating continuous glucose monitoring, dietary logs, and activity sensing. Over years, she tuned alert thresholds, meal suggestions, and exercise reminders, a process Shen et al. describe as customization via interactive alignment and adapted model learning [11].

After Ayumi dies, her grandson Kenji, 24, inherits access to her account.

Kenji is pre-diabetic and considers using his grandmother’s configuration as a starting point, assuming shared genetic risk. Yet the system remains calibrated to Ayumi’s 78-year-old physiology and disease stage. It begins producing recommendations optimized for an elderly woman with advanced diabetes rather than a young adult with early risk factors.

Alignment tension: Highly personalized alignment is often *non-transferable*. Years of co-adaptation produced an “alignment state” that is meaningful only within the Ayumi–AI relationship, and can become misleading or even dangerous when applied to a different body. While Shen et al. note customization as underexplored, this scenario sharpens the point: customization creates *relational artifacts* whose validity depends on the specific human–AI dyad that produced them. Inheritance breaks that dyad, leaving behind an optimized configuration without its original subject, and without a principled way to decide whether (or how) it should generalize.

5 Discussion

5.1 Posthumous Alignment as a Diagnostic Lens

These scenarios should not be treated as rare edge cases that can be fixed with small patches. We argue they are better understood as a *diagnostic lens*, a way to stress-test bidirectional alignment by asking a simple question: does the approach degrade gracefully when the human is no longer present?

Alignment has an expiration date. If alignment is dynamic and evolving [4, 11], then it also has temporal limits. A system

that is aligned with a user in 2024 may not be aligned with the same user in 2026, and it is even less likely to remain aligned after the user dies. Shen et al. emphasize that alignment must “adapt dynamically since human values and preferences change” [11]. We take that claim seriously and draw a corollary. If alignment requires ongoing change, systems must also be able to detect when the human partner has permanently exited the loop and treat that exit as a first-class alignment condition rather than an exception.

Relational alignment does not transfer cleanly. The Inherited Dashboard scenario (Section 4.3) highlights that deeply personalized alignment produces something closer to a relationship than a portable setting. It is situated, historically contingent, and bound to one specific human–AI pairing. This complicates the assumption implicit in Shen et al.’s discussion of “Customizing AI for Individuals or Groups” that personalized alignment states can be assembled from modular techniques and reused. Our scenarios suggest the opposite. Once alignment emerges through long-term co-adaptation, the resulting state is not merely a collection of tuned parameters. It is an interaction history that may not remain valid, safe, or meaningful when separated from the person who produced it.

Posthumous data requires alignment-aware governance. Current governance discussions often focus on privacy, consent, and ownership. We argue that governance must also incorporate alignment. Data collected under one alignment relationship can become misaligned when reused under another. This is especially salient for health sensing, where biometric data gathered for personal wellness can later be repurposed for population-scale model training [14]. If such repurposing is allowed, it should carry temporal and relational metadata. The system should encode not only what values were expressed, but also when, under which alignment relationship, and under what life circumstances those values were elicited.

5.2 Design Mechanisms for Temporal Boundaries

Moving from diagnosis to design, we sketch three mechanisms aimed at the temporal boundary of bidirectional alignment. These are not complete solutions. They are design provocations meant to show that the failures surfaced by our scenarios can be addressed through concrete HCI interventions.

Mechanism 1: Alignment Advance Directives. Medical advance directives allow patients to record treatment preferences for situations in which they can no longer communicate [13]. We propose an analogous mechanism for AI alignment. An Alignment Advance Directive would let users pre-specify how their aligned systems should behave after death. This is not equivalent to a simple deletion toggle. It must directly express the tensions our scenarios surface. For Maria (Section 4.1), a directive might permit certain forms of research reuse while prohibiting others, for example, “My biometric data may be used for aggregate population research, but my personalized health model should not be applied to individual patients.” For Jens (Section 4.2), it might require temporal provenance, for example, “Preserve the model version I last used for any posthumous reconstruction of my work. Do not apply later model updates to my interaction history.”

The design difficulty mirrors that of medical directives. People must specify preferences before they can anticipate future system capabilities. We suggest mitigating this by supporting *value-level* rather than *action-level* specification. Users express principles, such as “my data should support research but not commercial products,” and the system interprets those principles as capabilities evolve. To keep directives from becoming stale, the interface could prompt periodic review, for instance annually or after major system updates. This mirrors advance care planning, which is typically treated as an ongoing process rather than a one-time form [13].

Mechanism 2: Alignment Sunset Protocols. Instead of maintaining alignment indefinitely after death, systems could implement structured sunset protocols that transition data and models through explicit phases. We sketch three illustrative phases. First, *preservation* (0–6 months after death), in which the system freezes the alignment state at the last interaction, prevents retroactive application of new models to the deceased user’s history, and provides access to designated inheritors or researchers subject to the user’s directive. Second, *transition* (6–24 months), in which the system begins separating the user’s raw data from the personalized model derived from it. Aggregate and anonymized patterns may be retained for population-level research, but individualized models that encode personal routines and preferences are deprecated. Designated contacts receive explicit notice that the system is entering transition. Third, *dissolution* (24+ months), in which the personalized alignment model is retired entirely, and any retained data exists only in anonymized, aggregate form. At that point, the system stops presenting itself as if it remains in an active alignment relationship with the deceased.

The specific timelines are placeholders. The central claim is the principle. Alignment relationships should have designed endings, not indefinite persistence.

Mechanism 3: Inheritor Onboarding Interfaces. The Inherited Dashboard scenario (Section 4.3) shows that inheritors can encounter highly personalized systems without understanding the relationship that produced them. We propose an inheritor onboarding interface that makes alignment history explicit. When a new user accesses a system previously aligned to a deceased person, the interface would present an *alignment provenance summary*. This would describe, in plain language, who the system was aligned to, for how long, what personalization occurred, and why that state may not be appropriate for a new user. For Kenji inheriting Ayumi’s dashboard, the summary might state, “This system was personalized over six years for a 78-year-old woman with Type 2 diabetes. Alert thresholds, meal suggestions, and activity targets were calibrated to her physiology and disease progression. These settings may be medically inappropriate for other users.”

Crucially, the interface should not reduce the problem to a single “reset” button. Resetting discards potentially useful structure without explaining what made the inherited state risky. Instead, onboarding should support a guided *realignment* process. It would distinguish which elements may be transferred safely, such as dashboard layout; which are dangerous to reuse, such as medication-related thresholds; and which require expert input, such as clinically appropriate target ranges. This mechanism speaks directly to Shen et al.’s underexplored dimension of “Education and Training Human”. The goal is not general AI literacy, but rather to help users

understand how *this system's alignment history* shapes its recommendations and failure modes.

5.3 Extending the Bidirectional Framework

Shen et al.'s Bidirectional Human-AI Alignment framework [11] provides a valuable map of alignment research. We propose adding temporal boundary conditions as an explicit dimension. Each of their four research questions can be supplemented with a corresponding boundary question. What happens to value specification when the human source is no longer available? How should integrated values be handled when their origin is a deceased user? How do cognitive adjustments persist or decay when the AI counterpart changes after the human is gone? What does adaptive behavior mean for survivors and inheritors of an alignment relationship? The three mechanisms above, Alignment Advance Directives, Alignment Sunset Protocols, and Inheritor Onboarding, are initial responses that we offer as concrete prompts for workshop discussion.

6 Conclusion

This position paper contributes to the themes of *Value-Centered Alignment Objectives*, *Developing Interfaces and Interactions for Alignment*, and *Dynamic Co-Evolution of Human-AI Futures*. We frame posthumous alignment as a productive provocation, a boundary condition that forces the community to clarify what alignment means when core assumptions, mutual adaptation, ongoing consent, and a living human-in-the-loop no longer hold. By grounding our analysis in the Bidirectional Human-AI Alignment framework [11] and moving from scenario-based diagnosis to concrete design sketches, we aim to make the contribution actionable for both HCI and AI alignment communities. The three scenarios, The Silent Insole, The Evolving Assistant, and The Inherited Dashboard, are intended as discussion artifacts. The three mechanisms, Alignment Advance Directives, Alignment Sunset Protocols, and Inheritor Onboarding, are intended as starting points for collaborative design. We invite critical engagement with whether these mechanisms are sufficient, whether they introduce new value tensions, and what additional design patterns may be needed for alignment approaches that take human mortality seriously.

References

- [1] Jed R. Brubaker, Gillian R. Hayes, and Paul Dourish. 2013. Beyond the Grave: Facebook as a Site for the Expansion of Death and Mourning. *The Information Society* 29, 3 (2013), 152–163. doi:10.1080/01972243.2013.777300
- [2] Micah Carroll, Davis Foote, Anand Siththaranjan, Stuart Russell, and Anca Dragan. 2024. AI Alignment with Changing and Influenceable Reward Functions. In *ICLR 2024 Workshop: How Far Are We From AGI*. <https://openreview.net/forum?id=aC8D55CfeU>
- [3] Batya Friedman and David G. Hendry. 2019. *Value Sensitive Design: Shaping Technology with Moral Imagination*. MIT Press.
- [4] Iason Gabriel. 2020. Artificial Intelligence, Values, and Alignment. *Minds and Machines* 30, 3 (2020), 411–437.
- [5] Edina Harbinja. 2017. Legal Aspects of Transmission of Digital Assets on Death. In *Digital Legacy and Interaction*. Springer, 139–166.
- [6] Victor Kaptelinin and Bonnie A. Nardi. 2006. *Acting with Technology: Activity Theory and Interaction Design*. MIT Press.
- [7] Michael Massimi and Ronald M. Baecker. 2010. A Death in the Family: Opportunities for Designing Technologies for the Bereaved. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 1821–1830.
- [8] Carl Ohman and Luciano Floridi. 2017. The Political Economy of Death in the Age of Information: A Critical Approach to the Digital Afterlife Industry. *Minds and Machines* 27, 4 (2017), 639–662.
- [9] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training Language Models to Follow Instructions with Human Feedback. In *Advances in Neural Information Processing Systems*, Vol. 35. 27730–27744.
- [10] Shalom H. Schwartz. 2012. An Overview of the Schwartz Theory of Basic Values. *Online Readings in Psychology and Culture* 2, 1 (2012), 11.
- [11] Hua Shen, Tiffany Kneareem, Reshmi Ghosh, Kenan Alkiek, Kundan Krishna, Yachuan Liu, Savvas Petridis, Yi-Hao Peng, Li Qiwei, Chenglei Si, Yutong Xie, Jeffrey P. Bigham, Frank Bentley, Joyce Chai, Zachary Chase Lipton, Qiaozhu Mei, Michael Terry, Diyi Yang, Meredith Ringel Morris, Paul Resnick, and David Jurgens. 2025. Position: Towards Bidirectional Human-AI Alignment. <https://openreview.net/forum?id=PgA9rZoMY8>
- [12] Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, et al. 2024. A Roadmap to Pluralistic Alignment. *arXiv preprint arXiv:2402.05070* (2024).
- [13] Rebecca L. Sudore, Hillary D. Lum, John J. You, Laura C. Hanson, Diane E. Meier, Steven Z. Pantilat, Daniel D. Matlock, Judith A.C. Rietjens, Ida J. Korfage, Christine S. Ritchie, et al. 2017. Defining Advance Care Planning for Adults: A Consensus Definition from a Multidisciplinary Delphi Panel. *Journal of Pain and Symptom Management* 53, 5 (2017), 821–832.
- [14] Shoshana Zuboff. 2019. *The Age of Surveillance Capitalism*. PublicAffairs.