ROBOTIC COLLECTIVE MEMORY

MICHAL SHUR-OFRY* **& GUY PESSACH****

ABSTRACT

The various ways in which robots and AI will affect our future society are at the center of scholarly attention. This Commentary, conversely, concentrates on their possible impact on humanity's past, or more accurately, on the ways societies will remember their joint past. We focus on the emerging use of technologies that combine AI, cutting-edge visualization techniques, and social robots, in order to store and communicate recollections of the past in an interactive human-like manner. We explore the use of these technologies by remembrance institutions and their potential impact on collective memory. Taking a close look at the case study of NDT (New Dimensions in Testimony)—a project that uses 'virtual witnesses' to convey memories from the Holocaust and other mass atrocities—we highlight the significant value, and the potential vulnerabilities, of this new mode of memory construction.

Against this background, we propose a novel concept of memory fiduciaries that can form the basis for a policy framework for robotic collective memory. Drawing on Jack Balkin's concept of 'information fiduciaries' on the one hand, and on studies of collective memory on the other, we explain the nature of and the justifications for memory fiduciaries. We then demonstrate, in broad strokes, the potential implications of this new conceptualization for various questions pertaining to collective memory constructed by AI and robots. By so doing, this Commentary aims to start a conversation on the policies that would allow algorithmic collective memory to fulfill its potential, while minimizing its social costs. On a more general level, it brings to the fore a series of important policy questions pertaining to the intersection of new technologies and intergenerational collective memory.

* Hebrew University of Jerusalem Law Faculty.

** Hebrew University of Jerusalem Law Faculty. We thank Ryan Abbot, James Grimmelmann, Michal Lavi, Mark Lemley, Yafit Lev-Aretz, Tomer Kenneth, Gideon Parchomovsky, Katherine Strandburg, Ofer Tur-Sinai, Ari Waldman, Steven Wilf, as well as the participants of the Privacy Research Group at NYU Information Law Institute, for valuable comments and illuminating discussions. Anat Kahana and Uria Beeri provided excellent research assistance. This research was supported by the ISF, Grant Number 1342/17.

TABLE OF CONTENTS

INTRODUCTION, OR: MEET PINCHAS GUTTER

Pinchas Gutter, a Holocaust survivor, is sitting in a room full of students. "My name is Pinchas Gutter," he begins, "I will answer any questions you might have for me." A boy raises his hand. "How old were you when the War ended?", he asks. "I was between the ages thirteen and fourteen when the War ended. In 1945," Gutter answers. A girl asks: "Do you remember any songs from your youth?" Gutter smiles. "This is a lullaby that my mother used to sing to me, and I still remember it. It's in Polish." Still smiling, he starts singing. His audience is fascinated, only the real Pinchas Gutter is not in the room. The conversation takes place with a virtual Pinchas Gutter—a hologram-like image, backed by sophisticated software.[1] The system integrates advanced display technologies, complicated natural language processing AI, and a database of pre-recorded video interviews conducted with Gutter himself.[2] Their combination allows the 'virtual Gutter' to identify the audience's questions, match the most relevant response from the pre-existing database, and present the answer, as originally delivered by Gutter, in what simulates a human conversational interaction.[3]

1. A video of the above-quoted discussion is available on YouTube. ICT Vision & Graphics Lab, *New Dimensions in Testimony - USC ICT and SFI - Classroom Concept*, YOUTUBE (Feb. 8, 2013), https://www.youtube.com/watch?v=AnF630tCiEk.

2. A description of the technology can be found on the project's website, *Dimensions in Testimony*, USC SHOAH FOUND., https://sfi.usc.edu/collections/holocaust/ndt [https://perma.cc/D526-SVP9], and in Part II *infra*.

3. *Dimensions in Testimony*, *supra* note 2; *New Dimensions Body Text*, USC SHOAH FOUND., https://sfi.usc.edu/content/new-dimensions-body-text [https://perma.cc/CZ8Z-ZHA3].

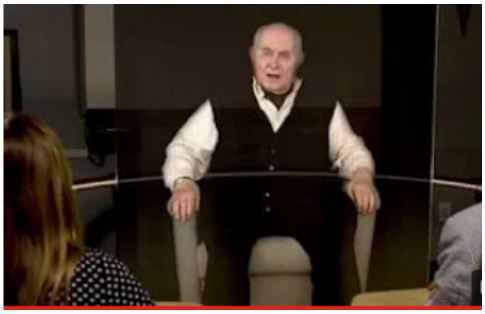

Figure 1: *The Virtual Pinchas Gutter in a Classroom*[4]

The virtual Gutter is part of New Dimensions in Testimony (NDT)—a pioneering project of the USC Shoah Foundation that enables people to have conversations with pre-recorded videos of Holocaust survivors and other witnesses to genocide.[5] In a sense, these virtual witnesses are part of a growing phenomenon of AI-based social robots—robots that are engineered to engage with humans in a social-like manner, exercising learning, communication, and adaptive software capabilities.[6] With the development of machine learning and visualization techniques, the use of AI and social robots is expanding, and their impact on future human lives has become the subject of intense law and policy discussion.[7]

Yet, what is largely missing from this conversation, and what is striking in the case of the virtual Pinchas Gutter, is the use of AI and social robots in a way that affects humanity's *past*, or more accurately our collective memory of the past. This Commentary uses the NDT project as a starting

4. *See* ICT Vision & Graphics Lab, *supra* note 1.

5. *Id.*

6. In this Commentary, we use the term "robots" in a rather elaborated way, to include not only robots embodied in a material object, but also other AI and machine learning agents that interact with their 'end-users.' We also use the terms "robotic," "virtual," and "algorithmic" memory agents interchangeably. While certain distinctions exist, they are immaterial for the purpose of the present discussion. *See infra* notes 48–52 and accompanying text. For the development of the social robots concept, see Cynthia Breazeal, *Towards Sociable Robots*, 42 ROBOTICS & AUTONOMOUS SYSTEMS 167, 174 (2003) (discussing categories of social robots and arguing that "endowing a robot with social skills and capabilities has benefits far beyond the interface value for the person who interacts with it"). For examples of various social robots and a discussion of their prominent attributes, see *infra* Part II.

7. *See* Frank Pasquale & Arthur J. Cockfield, *Beyond Instrumentalism: A Substantivist Perspective on Law, Technology, and the Digital Persona*, 2018 MICH. ST. L. REV. 821, 842–44 (referring to a long line of works that identify how AI and digital technology transform "human experience, identity, and aims"). For discussions of the impact of social robots, see, for example, Kate Darling, *Extending Legal Protection to Social Robots: The Effects of Anthropomorphism, Empathy, and Violent Behavior Towards Robotic Objects*, *in* ROBOT LAW 213 (Ryan Calo, A. Michael Froomkin & Ian Kerr eds., 2016) [hereinafter Darling, *Social Robots*]; Kate Darling, "*Who's Johnny?" Anthropomorphic Framing in Human-Robot Interaction, Integration, and Policy*, *in* ROBOT ETHICS 2.0: FROM AUTONOMOUS CARS TO ARTIFICIAL INTELLIGENCE 173 (Patrick Lin, Ryan Jenkins & Keith Abney eds., 2017) [hereinafter Darling, *Johnny*] (describing the phenomenon of social robots and discussing related ethical aspects); *infra* notes 53–60.

point for a broader discussion of the policy questions pertaining to the interface of AI, collective memory, and the law. How will societies remember the joint past in an era of virtual memory agents? Should the law regulate the use of AI and robots in ways that affect collective memory, and if so, what would be an appropriate policy framework? These questions are at the center of our inquiry. While our point of departure is the case study of NDT's virtual witnesses, our analysis applies to a broader range of cases where collective memory is mediated through AI-based technologies that possess interactive-communicative skills and perceived human-like authenticity.[8]

Part I of this Commentary begins with a brief introduction of collective memory, a concept that is the subject of burgeoning interdisciplinary literature, yet is still largely new to legal analysis. We briefly explain the notion of collective memory, its social value for the construction of collective and individual identities, and the multiple ways that affect its formation.

Part II takes a closer look at the emerging use of robotic memory agents in the construction of collective memory. Relying on interdisciplinary studies, we show that this new medium carries great promise; it allows for interactions that feel natural, encourages trust and empathy, and can help bridge temporal gaps. It may also be particularly important for overcoming or mitigating 'problems of representation' that exist in cases of genocide or other extreme events.[9] Following this discussion, we proceed to explore potential concerns entailed in this new medium of memory construction, identifying two primary challenges: First, the use of AI-based memory agents inevitably involves editorial choices that may not be transparent to their 'end-users.' While such choices are an unavoidable part of each medium that provides information, the traits of robotic memory agents might make these choices particularly invisible. Secondly, these 'modes of memory' are more susceptible to hacking, manipulation by third parties, and other vulnerabilities in comparison to more traditional media that affect memory construction. These concerns are particularly pronounced since this new medium, by its nature, evokes feelings of trust and reliance on part of its 'users.'

Against this backdrop, Part III of this Commentary explores the potential policy responses to these developments. Relying on socio-cultural studies of collective memory and building on the concept of 'information

8. As we explain in Parts II and III *infra*, our analysis also encompasses algorithmic memory agents that are based on verbal interactions without a visual interface.

9. For the term 'problem of representation,' see *infra* note 25 and accompanying text.

fiduciaries' developed by Jack Balkin,[10] we introduce a new concept of 'memory fiduciaries.' We explain the nature of and the justifications for memory fiduciaries, and demonstrate, in broad strokes, the potential implications of this new conceptualization for various questions pertaining to collective memory constructed by robots. Our purpose is neither to exhaust the discussion, nor to present a comprehensive 'menu' of legal solutions. Rather we aim to start a policy discussion about the important questions that are at the interface of collective memory, new technologies, and the law.

I. A BRIEF PRIMER TO COLLECTIVE MEMORY

The notion of collective memory refers to the joint recollection of the past by societies, communities, nations, and additional groups with elements of a joint identity.[11] Largely attributed to sociologist Maurice Halbwachs, the modern concept of collective memory relies on the understanding that memories are, to an extent, a product of social construction.[12] In other words, our memories are not merely the sum of our own individual experiences, but are constructed in part by the groups to which we belong, be they nations, religious groups, minority groups, kinship networks, or other communities. To illustrate, many of us would say, in everyday parlance, that we remember the first human landing on the moon, the Kennedy assassination, or the Holocaust, although we did not personally experience these events and may not have even been born when they occurred.[13] Thus, the focus of collective memory is not on the cognitive processes of individual memory formation and retrieval, but rather on the social processes and elements that shape our memories as groups.[14]

During the past few decades, the study of collective memory has rapidly developed into a burgeoning, multi-disciplinary field, integrating insights from sociology, history, anthropology, communication studies, and

10. Jack M. Balkin, *Information Fiduciaries and the First Amendment*, 49 U.C. DAVIS L. REV. 1183 (2016) [hereinafter Balkin, *Information Fiduciaries*].

11. *See, e.g.*, Jeffrey K. Olick & Joyce Robbins, *Social Memory Studies: From "Collective Memory" to the Historical Sociology of Mnemonic Practices*, 24 ANN. REV. SOC. 105, 106 (1998); Jeffrey K. Olick, Vered Vinitzky-Seroussi & Daniel Levy, *Introduction* to THE COLLECTIVE MEMORY READER 3, 16–22 (Jeffrey K. Olick, Vered Vinitzky-Seroussi & Daniel Levy eds., 2011); Jeffrey K. Olick, *Collective Memory: The Two Cultures*, 17 SOC. THEORY 333, 334–35 (1999).

12. *See* MAURICE HALBWACHS, ON COLLECTIVE MEMORY 37–40 (Lewis A. Coser ed. and trans., 1992).

13. *See, e.g.*, Eviatar Zerubavel, *Social Memories: Steps to a Sociology of the Past*, 19 QUALITATIVE SOC. 283, 289–90 (1996) (explaining that people share the memories of the groups to which they belong, even when they did not individually experience them).

14. *See, e.g.*, Olick & Robbins, *supra* note 11, at 106–08 (discussing the development of the perception of collective memory as a social construction of the past).

additional areas.[15] This scholarship recognizes that collective memory forms the connection between groups and their past.[16] It is thus necessary for narrating the life-stories of nations and communities, and constitutes a vital part of their collective identities.[17] Moreover, when the relevant groups are minorities or groups that were exposed to atrocities and persecution, collective memory is perceived as a means of empowerment and restoration.[18] Studies further instruct that collective memory is important for the formation of individual identity as well, since the social and cultural groups of which we are a part deeply influence our sense of self and identity.[19]

Although often relying on historical accounts, collective memory is not synonymous with history. Since it is a product of social construct, it has subjective and normative dimensions, and can more easily encompass a multiplicity of voices and meanings.[20] Indeed, the same historical event—for example, the atomic bomb on Hiroshima—can play an entirely different role in the collective memory of different groups.[21]

---

15. *See, e.g.*, Olick & Robbins, *supra* note 11, at 106.

16. *See, e.g.*, Amos Funkenstein, *Collective Memory and Historical Consciousness*, 1 HIST. & MEMORY 5, 5 (1989) ("[W]ithout memory of the past there is no history, in the sense of the events that are meaningful to the collective, events experienced by a collective that is aware of them.").

17. *See, e.g.*, Olick, *supra* note 11, at 333 ("Collective memory . . . often plays an important role in politics and society."); Zerubavel, *supra* note 13, at 290 ("[B]eing social presupposes the ability to experience events that had happened to groups and communities to which we belong long before we joined them as if they were part of our own past . . . ."); Barbie Zelizer, *Reading the Past Against the Grain: The Shape of Memory Studies*, 12 CRITICAL STUD. MASS COMM. 214, 226–28 (1995); Jan Assmann, *Collective Memory and Cultural Identity*, 65 NEW GERMAN CRITIQUE 125, 126 (John Czaplicka trans., 1995); IWONA IRWIN-ZARECKA, FRAMES OF REMEMBRANCE: THE DYNAMICS OF COLLECTIVE MEMORY 47–57 (1994).

18. *See, e.g.*, Sharon K. Hom & Eric K. Yamamoto, *Collective Memory, History, and Social Justice*, 47 UCLA. L. REV. 1747, 1758 (2000) ("Collective memory not only vivifies a group's past, it also reconstructs it and thereby situates a group in relation to others in a power hierarchy."); Sara Jones, *"Simply a Little Piece of GDR History"?: The Role of Memorialization in Post-Socialist Transitional Justice in Germany*, 27 HIST. & MEMORY 154 (2015) (considering the role that collective memorialization plays in transitional justice).

19. *See, e.g.*, Susan A. Crane, *Writing the Individual Back into Collective Memory*, 102 AM. HIST. REV. 1372, 1381–83 (1997) (discussing the relations between individuals and collective memory); Assmann, *supra* note 17, at 127; W. James Booth, *The Work of Memory: Time, Identity, and Justice*, 75 SOC. RES. 237 (2008) (discussing the value of collective memory for the formation of individual identity).

20. *See* Olick & Robbins, *supra* note 11, at 110 (discussing the relations between social memory studies and historiography); Funkenstein, *supra* note 16, at 5 (referring to Hegel's conception of history and to the distinction between memory and history); Steven Knapp, *Collective Memory and the Actual Past*, 26 REPRESENTATIONS 123, 141 (1989) (explaining that "shared values are likely to be connected to the narratives preserved by collective memories").

21. *See, e.g.*, Stefanie Fishel, *Remembering Nukes: Collective Memories and Countering State History*, 1 CRITICAL MIL. STUD. 131, 136–37, 141 (2015) (comparing the different memorialization of the use of nuclear weapons in Japan and the United States, and describing how in Japan the bomb was commemorated as a sacrifice for peace, while in the United States, a Smithsonian exhibition that meant to convey a message that "a mission using nuclear weapons against human beings should not be glorified" was denounced by the Senate and eventually cancelled).

Relatedly, multiple sources affect the formation of collective memory. A non-exhaustive list includes historical and documentary materials, formal and informal studies, media coverage, visits to physical sites, as well as community rituals and witness testimonies.[22] From the perspective of collective memory construction, the latter are particularly important. The reason is that witnesses are able to convey stories and experiences in a direct, non-mediated way that "save[s] the imagination from abstraction."[23] While people may perceive archival materials as reflections of distant events that are "in the past," witness testimonies create an intimate effect that bridges this gap and allows the distance to disappear.[24] This is particularly important for the collective memory of traumatic and radical events, such as genocide, where the extremity and magnitude of the event makes it particularly difficult to grasp by ordinary means of documentation and storytelling.[25] To illustrate, Holocaust research indicates that the testimonies of thousands of Holocaust survivors, describing "the fate of one person and then another, of one family and then another," helped transform the abstract concept of "six million" Jewish Holocaust victims into something more concrete and comprehensible.[26] To use the words of director Claude Lanzmann, "There was an absolute break between the bookish knowledge I had acquired and what these people told me."[27] This understanding initiated the formation of archives comprised of large collections of witness testimonies, such as the Fortunoff archive at Yale

22. *See, e.g.*, Olick & Robbins, *supra* note 11, at 106–08 (describing various sources that play a role in social memory construction); HALBWACHS, *supra* note 12, at 52–53 (discussing the spatial aspects of collective memory); Geoffrey H. Hartman, *Learning from Survivors: The Yale Testimony Project*, 9 HOLOCAUST & GENOCIDE STUD. 192 (1995) (analyzing the significance of witness testimonies).

23. Hartman, *supra* note 22, at 192.

24. *Id.* at 198.

25. In the context of the Holocaust, this phenomenon is often referred to as "the problem of representation," see Saul Friedlander, *Introduction* to PROBING THE LIMITS OF REPRESENTATION: NAZISM AND THE "FINAL SOLUTION" 1, 3 (Saul Friedlander ed., 1992); Geoffrey H. Hartman, *Introduction: Darkness Visible*, *in* HOLOCAUST REMEMBRANCE: THE SHAPES OF MEMORY 1, 2, 5–6 (Geoffrey H. Hartman ed., 1994) (discussing the difficulties to represent the Holocaust by traditional means and modes of representation).

26. Hartman, *supra* note 22, at 195 (referring to video testimonies of Holocaust survivors that comprise the Yale Testimony Project).

27. Hartman, *supra* note 22, at 203 (quoting Claude Lanzmann, *Le Lieu et la Parole*, *in* AU SUJET DE SHOAH: LE FILM DE CLAUDE LANZMANN 293, 294 (Michel Deguy ed., 1990) (referring to the renowned film "*Shoah*")).

University,[28] and the archive of the USC Shoah Foundation.[29] The latter is also the source of the NDT technology to which we turn shortly.

These collections highlight a more general point: witness testimonies, like other sources that comprise collective memory, are often collected and mediated to the public through 'remembrance institutions'—entities such as archives, libraries, and museums, which select, document, preserve, and provide access to various materials and artifacts.[30] Obviously, remembrance institutions are not the sole entities that mediate these materials and affect collective memory. Multiple other sources play a role in its construction, including social media, fiction films, documentaries, popular press, or education systems. Yet, the role of remembrance institutions is prominent, especially with respect to events that produce abundant piecemeal materials. Consider, for example, the American Civil War.[31] Each piece alone is unlikely to have sufficient market demand, and therefore has no real prospects of being distributed through commercial market channels. Yet, the pieces' collection together by a remembrance institution enables us to draw a 'big picture' that is greater than the sum of its components, and has a substantial impact on collective memory.[32]

As the discussion below demonstrates, the role of these institutions is becoming all the more significant when artificial intelligence and robots are involved in the mediation of memories.[33] The next Part thus returns to the virtual Gutter, and explores the potential effect of these technological developments on the formation of collective memory.

28. The Fortunoff archive began as a grassroots enterprise in New Haven, and currently holds more than 4,400 testimonies comprising 12,000 recorded hours of videotape. *See Fortunoff Video Archive for Holocaust Testimonies*, YALE, https://fortunoff.library.yale.edu/ [https://perma.cc/DHG6-E799].

29. The USC Shoah Foundation's Archive currently has more than 55,000 video testimonies. Most of those testimonies were given by Holocaust survivors, but the archive has expanded to include testimonies from other cases of genocide and mass atrocities, including the 1994 Rwandan Genocide, the 1937 Nanjing Massacre, the Armenian Genocide, and the Guatemalan Genocide. *See About Us*, USC SHOAH FOUND., https://sfi.usc.edu/about [https://perma.cc/RXX2-N43H].

30. Guy Pessach, *[Networked] Memory Institutions: Social Remembering, Privatization and Its Discontents*, 26 CARDOZO ARTS & ENT. L.J. 71, 73 (2008); Guy Pessach & Michal Shur-Ofry, *Copyright and the Holocaust*, 30 YALE J.L. & HUMAN. 121, 136–37, 158 (2018) (discussing the role of remembrance institutions with respect to the Holocaust's collective memory).

31. The abundant piecemeal materials from the Civil War include, among others, thousands of photos, war maps, cartoons, battle chronicles, newspaper articles, rosters of soldiers, and additional official documents. *See, e.g.*, *The Civil War in America*, LIBR. CONGRESS, https://www.loc.gov/exhibits/civil-war-in-america/learn-more.html [https://perma.cc/39FT-JRSS].

32. *Cf.* Pessach & Shur-Ofry, *supra* note 30, at 168–69.

33. *See infra* notes 92–96 and accompanying text.

II. COLLECTIVE MEMORY IN THE AGE OF AI

A. Virtual Memory Agents and Social Robots

The virtual Pinchas Gutter, along with additional 'virtual witnesses' that form part of the NDT project, reflect the recognition of the significance of witness testimonies, and the realization that at a certain point, live witnesses of the Holocaust and other genocides will no longer be available.[34] The introduction of the technology seems to have been accompanied by a remarkable degree of reflection and self-awareness on part of the institutions involved, making NDT a particularly apt case study for our purposes.[35]

According to information supplied by the NDT project, the interaction with each 'virtual witness' relies on a database of answers and recollections of the real survivors, who were filmed answering thousands of questions in a long and detailed interview process.[36] The database connects with a natural language processing software with learning capabilities, which is able to "understand" the questions people ask the virtual witness. Based on the recognition of similarities between word patterns in the end-users' questions and the answers given by the original survivor, the software selects the most relevant answer from the database.[37] The data is then captured and played back verbatim, as delivered by the survivor.[38] Thus, the technology enables the virtual witness to seamlessly answer the question posed, using the original answers of the actual survivor. [39]

---

34. *See New Dimensions Body Text*, *supra* note 3 ("Years from now, long after the last [Holocaust] survivor has left us, Dimensions in Testimony will be able to provide a valuable opportunity to engage with a survivor and ask them questions directly . . . .").

35. For example, the project was extensively discussed in a conference titled "Digital Approaches to Genocide Studies" held by the USC Shoah Foundation in 2017. *See Institute News: Scholars Consider Ethics, Possibilities, and Critiques of New Dimensions in Testimony at Digital Approaches to Genocide Studies Conference*, USC SHOAH FOUND. (Nov. 21, 2017), https://sfi.usc.edu/news/2017/11/20081-scholars-consider-ethics-possibilities-and-critiques-new-dimensions-testimony [https://perma.cc/XAP9-J7ZR] [hereinafter *Scholars Consider*].

36. *Technology in Service to Humanity*, USC SHOAH FOUND. (Nov. 2017), https://sfi.usc.edu/sites/default/files/.../dit_one_sheet_holocaust_20181019_opt.pdf [https://perma.cc/4K8U-54WM] ("During the interview process, the survivor sits in the middle of a light stage beneath a half dome latticed with lights and more than 100 video cameras. Each subject answers as many as 2,000 questions that cover a vast range of subjects.").

37. *New Dimensions in Testimony*, USC SHOAH FOUND. (Apr. 2017), https://www.ilholocaustmuseum.org/wp-content/uploads/2017/09/New-Dimensions-in-Testimony-one-sheet.pdf [https://perma.cc/Q8ZP-ZQP8] (describing the operation of the natural language processing software).

38. *Technology in Service to Humanity*, *supra* note 36 ("Using natural-language technology, the program matches questions with the survivor's most relevant response."); *New Dimensions in Testimony*, *supra* note 37 ("Whether people ask, "Where were you born?", "Do you believe in God?", "How did you survive?", data is captured and processed into video segments that can be played back verbatim, precisely as the survivors delivered them.").

39. *See Technology in Service to Humanity*, *supra* note 36; *New Dimensions in Testimony*, USC INST. FOR CREATIVE TECHS., http://ict.usc.edu/wp-content/uploads/overviews/New%20Dimensions%2

The sophisticated visualization of the virtual witness, although not strictly a hologram, further amplifies the interactive, natural conversation experience.[40] Future development of the technology will further enhance this feeling of natural interaction. For example, over time the algorithms underlying the virtual witnesses will learn to respond to vocal cues signifying age, and select the answers accordingly.[41] Likewise, advances in visualization techniques will allow the display of the witnesses in three, rather than two, dimensions.[42]

Virtual witnesses are certainly a disruptive technology in the field of collective memory. Yet, this development should not be viewed in isolation. In order to normatively evaluate it, one should locate it against a broader field of technologies that aim to create a natural interaction experience between humans and algorithms. To use some famous examples, the newly introduced Google Duplex technology can conduct natural conversations in order to carry out specific tasks, such as scheduling appointments over the phone.[43] By mimicking ordinary interaction, including the incorporation of speech disfluencies—"hmm"s and "uh"s—the system allows people to speak normally, without having to adapt to a machine.[44] Likewise, the USC Institute for Creative Technologies, which created the technologies underlying the NDT project, developed a number of virtual human characters with similar capabilities, for purposes such as training psychologists, treating soldiers who underwent traumatic experiences, or sparking interest in science and technology among young people.[45]

These technologies are paralleled by significant developments in the fields of visualization and imaging. To illustrate, scientists have recently created a photorealistic 'talking head' model of Barack Obama—an AI-based virtual video of Obama, who moves his (virtual) lips in synchronization with speech implanted by the researchers in a seemingly

---

0in%20Testimony_Overview.pdf [https://perma.cc/AP66-PFRR]; Eillie Anzilotti, *So We Never Forget, Holograms Will Keep Delivering First-Person Holocaust Survivor Testimony*, FAST COMPANY (June 20, 2017), https://www.fastcompany.com/40427922/so-we-never-forget-holograms-will-keep-delivering-first-person-holocaust-survivor-testimony [https://perma.cc/S42C-U7GH].

40. *Technology in Service to Humanity*, *supra* note 36; Anzilotti, *supra* note 39; *New Dimensions in Testimony*, *supra* note 37.

41. Anzilotti, *supra* note 39 (quoting one of the project's leaders).

42. *New Dimensions in Testimony*, *supra* note 37 ("Soon, visualization techniques in development will be able to display the survivor in three dimensions—no 3-D glasses required—to provide an experience as close as possible to face-to-face interaction.").

43. Yaniv Leviathan & Yossi Matias, *Google Duplex: An AI System for Accomplishing Real-World Tasks Over the Phone*, GOOGLE AI BLOG (May 8, 2018), https://ai.googleblog.com/2018/05/duplex-ai-system-for-natural-conversation.html [https://perma.cc/N8SM-BRX2].

44. *Id.*

45. *See All Prototypes*, USC INST. FOR CREATIVE TECHS., http://ict.usc.edu/prototypes/all/ [https://perma.cc/6GGU-Z4Z4].

natural way.[46] Further, the current state of three-dimensional holograms allows virtual participation in meetings and conferences, while bridging geographical distance and creating a more personal communication experience.[47]

More generally, these technologies can be viewed as part of the expanding use of social robots, namely AI that is engineered to engage with humans in a social-like manner, demonstrating adaptability, learning, and communication capabilities.[48] Examples include interactive toys, such as robotic dogs, dolls, and dinosaurs,[49] robots that function as personal assistants, and robotic companions that serve medical and social purposes.[50] Indeed, the virtual witnesses we focus on might not strictly qualify as 'robots,' since they may lack a physical embodiment.[51] Nevertheless, much like physical social robots, their essence lies in their ability to interact with humans in a sociable manner.[52] Therefore, viewing virtual memory agents through the prism of social robots allows us to better assess the benefits and challenges entailed in their use in contexts that affect collective memory. We turn to this analysis in the following section.

---

46. Supasorn Suwajanakorn, Steven M. Seitz & Ira Kemelmacher-Shlizerman, *Synthesizing Obama: Learning Lip Sync from Audio*, 36 TRANSACTIONS ON GRAPHICS 95 (2017).

47. *See, e.g.*, Elizabeth Gibney, *Physicists Create* Star Wars-*Style 3D Projections — Just Don't Call Them Holograms*, NATURE (Jan. 24, 2018), https://www.nature.com/articles/d41586-018-01125-y [https://perma.cc/RU8T-WZAK]; Jena McGregor, *'Star Wars' Meets the C-suite: This CEO's Hologram Is Beaming into Meetings*, WASH. POST (Apr. 13, 2016, 6:08 AM), https://www.washingtonpost.com/news/on-leadership/wp/2016/04/13/star-wars-meets-the-c-suite-why-this-ceos-hologram-is-getting-beamed-into-meetings/?noredirect=on (detailing the use of holograms in multi-national corporations and quoting Accenture's CEO: "I believe my hologram might be as good as me").

48. *See* Breazeal, *supra* note 6, at 167; Darling, *Social Robots*, *supra* note 7, at 213; Ari Ezra Waldman, *Safe Social Spaces*, 96 WASH. U. L. REV. 1537, 1560–62 (2019) (explaining that social robots display "social abilities, including communication, cooperation, and learning" and further discussing the entailed risks to privacy).

49. *See* sources cited *supra* note 48; *see also* Eldar Haber, *Toying with Privacy: Regulating the Internet of Toys*, 80 OHIO ST. L.J. 399, 401 (2019).

50. Darling, *Social Robots*, *supra* note 7, at 213.

51. *See* Ryan Calo, *Robotics and the Lessons of Cyberlaw*, 103 CALIF. L. REV. 513, 532 (2015) (listing physical embodiment as one of the attributes of social robots); Waldman, *supra* note 48, at 1561 (discussing the physical dimension of social robots). *But cf.* Neil M. Richards & William D. Smart, *How Should the Law Think About Robots?*, *in* ROBOT LAW, *supra* note 7, at 3, 5–6 (discussing the ambiguity in the definition of a robot); Mark A. Lemley & Bryan Casey, *You Might Be a Robot*, CORNELL L. REV. (forthcoming 2019), https://ssrn.com/abstract=3327602 (discussing the problems of defining robots and arguing that such ex ante definition may be impossible).

52. Therefore, as explained in the Introduction, in this Commentary we use the term "robots" in an elaborated way that is not restricted to physical robots. *Cf.* Jack M. Balkin, *The Three Laws of Robotics in the Age of Big Data*, 78 OHIO ST. L.J. 1217, 1219 (2017) [hereinafter Balkin, *Three Laws*] ("When I talk of robots, however, I will include not only robots—embodied material objects that interact with their environment—but also artificial intelligence agents and machine learning algorithms.").

B. Benefits and Challenges

One prominent attribute of social robots is the response they evoke from humans. Research indicates that robots' interactive behavior and ability to communicate and cooperate with people trigger anthropomorphism—a tendency to ascribe human qualities to the AI.[53] In other words, the verbal skills, the adaptability, and the seemingly-autonomous actions distinguish these systems in the human mind from mere machines, and make us react to them as if they were humans.[54] This tendency is intensified when the system appears human and animated, and uses facial expressions that we recognize and intuitively relate to, as is the case in the NDT project.[55] Interestingly, the effect subsists even when people are aware that they are interacting with a robot, and does not seem to disappear even for sophisticated users who are fully informed about the technology underlying the system.[56]

These traits shed light on the benefits of using robotic memory agents by remembrance institutions. As the case of Pinchas Gutter demonstrates, virtual witnesses can indeed connect with people in a way that elicits empathy and trust. The ability to interact, receive a response, form eye contact with the virtual witness, and follow his body language allows the users to form personal, intimate connection that is not possible when exposed to written or even video-taped testimonies.[57] While not equivalent to speaking to a natural person, this interactive mode of testimony creates an almost-natural conversation feeling.[58] As technology evolves, the sense

---

53. *See* Breazeal, *supra* note 6, at 168 (explaining that combining the robot's learning ability, creature-like behavior, and its "ability to communicate with, cooperate with, and learn from people makes it almost impossible for one to not anthropomorphize [it] (i.e., attribute human or animal-like qualities)").

54. *Id.*; Darling, *Social Robots*, *supra* note 7, at 218 ("[Social robots] elicit emotional reactions from people that are similar . . . to how we react to animals and to each other.").

55. Darling, *Social Robots*, *supra* note 7, at 218–19 (discussing the robot's animated look as a factor that affects anthropomorphism).

56. Matthias Scheutz, *The Inherent Dangers of Unidirectional Emotional Bonds Between Humans and Social Robots*, *in* ROBOT ETHICS: THE ETHICAL AND SOCIAL IMPLICATIONS OF ROBOTICS 205, 213–14 (Patrick Lin, Keith Abney & George A. Bekey eds., 2012) (describing this effect among roboticists from the MIT Media Lab); Darling, *Johnny*, *supra* note 7, at 173 ("Research shows that humans tend to anthropomorphize robotic technology, treating it as though it were alive, *even if we know better*." (emphasis added)).

57. *See Scholars Consider*, *supra* note 35 (quoting researcher Noah Shenker: "There was something incredibly ritualistic and quite moving about the encounters between the users of the testimony and Pinchas . . . . The nodding of the head as he listened, the eye contact that was maintained between most of the users and Pinchas . . . .").

58. *See, e.g.*, Christina Newland, *These VR Films Let Viewers Talk to Refugees and Holocaust Survivors*, VICE: MOTHERBOARD (June 17, 2016, 12:40 PM), https://www.vice.com/en_us/article/kb733e/vr-sheffield-doc-fest-talk-to-refugees-and-holocaust-survivors [https://perma.cc/GL65-LGXH] ("[The virtual Gutter] maintains eye contact and gives a real sense of naturalism, offering the spectator a personal connection with his life."); *Scholars Consider*, *supra* note 35 (quoting researcher Dan

of natural interaction will likely increase and further enhance the effectiveness of the testimony. Thus, virtual witnesses enable remembrance institutions to narrow the gap previously described between events that are 'in the past' and events that are socially relevant to our lives today.[59] Moreover, in the case of radical events, these technologies can significantly contribute to reducing problems of representation, making the abstract more concrete and perhaps a little more comprehensible.[60]

Alongside these significant virtues, there are, of course, challenges. Literature exploring the utilization of AI and social robots in various contemporary contexts often concentrates on the threats these technologies pose to their users' privacy.[61] Yet, in the context of collective memory construction, we believe the focal point of the discussion should be different.

A major challenge in our case lies in the fact that algorithmic memory agents are not neutral representations of past events, or even of witnesses' memories of those events. Rather, they inevitably involve, and reflect, a set of the editorial choices made by their creators. The NDT project, for example, is fraught with such decisions. Examples include selecting the live witnesses participating in the project; determining the location, length, and angles of filming; choosing the number, order, and nature of questions directed at the witnesses, as well as training the natural language processing AI that intermediates between the databases' contents and the users of the testimonies. Each such decision may influence the interactions between the users and the virtual witness, and as a result, bears significance for the construction of collective memory.[62]

Leopard: "Pinchas is a much more visual representation of a human, but there are stoppages, ... The consciousness is still in the realm of the uncanny").

59. Anzilotti, *supra* note 39 (quoting one of the NDT project's leaders: "We find that when a survivor speaks to a classroom or in the public domain, that impact that the meet-and-greet, the questions and answers, has on people and how we understand that history is significant"); *cf.* Noah Shenker, *Through the Lens of the Shoah: The Holocaust as a Paradigm for Documenting Genocide Testimonies*, 28 HIST. & MEMORY 141, 141–42 (2016) (describing the struggle of archives and museums to "preserve and circulate survivor testimonies of the Holocaust for future generations in ways that are socially relevant to those who will have had no exposure to living witnesses").

60. *See supra* notes 23–29 and accompanying text.

61. *See, e.g.*, M. Ryan Calo, *People Can Be So Fake: A New Dimension to Privacy and Technology Scholarship*, 114 PENN ST. L. REV. 809 (2010) (exploring the implications of technologies that imitate people for traditional privacy values); Waldman, *supra* note 48 (describing the potential hazards of technologies that mediate social interaction for users' privacy and safety); Haber, *supra* note 49 (discussing privacy concerns related to interactive toys).

62. *Cf.* Michal Shur-Ofry, *Databases and Dynamism*, 44 U. MICH. J.L. REFORM 315, 325–28 (2011) (observing, with respect to databases: "By determining the scope of information included in the database and the manner in which that information is accessed and retrieved, selections and arrangements contextualize database content and influence the manner in which that content is understood and interpreted by users").

Notably, the NDT case study is not an extreme example from the perspective of editorial discretion. According to the information they released, the project's creators seem to have taken considerable effort to minimize the inevitable gap between the original testimonies and the virtual witnesses, through a variety of means, including the use of the witnesses' original answers as output.[63] Nevertheless, some discrepancies inevitably remain. Thus, research indicates that the new medium of virtual witnesses yields testimony that is "led by the users," namely comprised of answers to users' questions, in comparison to testimonies of actual witnesses telling their stories, which are much more driven by the witnesses themselves.[64]

In addition, one can easily envisage other uses of AI for collective memory construction that would result in a greater discrepancy between the underlying materials and the AI output. Imagine, for example, an AI that integrates a large collection of testimonies from the Vietnam War, thus creating the 'ultimate witness'—one that delivers an integrated testimony about the War—or a 'virtual Abraham Lincoln'—one that relies on the 40,550 'Lincoln papers' stored at the Library of Congress to answer people's questions.[65] One can also imagine less benign cases, where interested parties may use this new technology to advance 'alternative' or biased narratives, while concealing their editorial discretion.

Importantly, editorial decisions that result in a gap between the raw information and its representation are not confined to AI or even to digital technology. Every medium that conveys information necessarily involves the judgment of its creators regarding the meaning and importance of that information.[66] Archives, databases, exhibitions, video collections, and other modes of memory construction always entail selection, discretion, and

63. *See* the technology description, *supra* notes 35–42 and accompanying text.

64. *See Scholars Consider*, *supra* note 35 (quoting Noah Shenker: "The experience . . . now focused on the user—the agency of the survivor was moved to user-driven imperatives. . . . Pinchas no longer speaks to listeners from start to finish, but we must ask questions to trigger sporadic narratives"); Noah Shenker & Dan Leopard, Presentation at the Memory Studies Association Conference: Pinchas Gutter: The Virtual Holocaust Survivor as Embodied Archive (Dec. 2017) (referring to the virtual testimony as "testimony on demand").

65. For information about the Lincoln Papers Collection, see *Abraham Lincoln Papers at the Library of Congress*, LIBR. OF CONGRESS, https://www.loc.gov/collections/abraham-lincoln-papers/about-this-collection/ [https://perma.cc/8R2G-ADVC]. While the example is hypothetical, technologies that attempt to construct a virtual Lincoln have existed since 1964, when Disney introduced its "Audio-Animatronics" version of Lincoln at the New York World's Fair. *See The Disneyland Story Presenting Great Moments with Mr. Lincoln*, DISNEYLAND, https://disneyland.disney.go.com/attractions/disneyland/disneyland-story?int_cmp=SOC-intDPFY11Q3NewTechBornAtDLR21-04-11@0004 [https://perma.cc/E3SV-X87Y] (mentioning that the version "was so life-like that *National Geographic* magazine called the figure 'alarming' in its realism"). For a current, more advanced version of the animated Lincoln, see Bus. Insider, *This Abraham Lincoln Animatronic Is So Life-Like You'll Feel like You're in 1863*, YOUTUBE (Dec. 19, 2017), https://www.youtube.com/watch?v=f6hEgDDRYds.

66. *Cf.* Niva Elkin-Koren, *Cyberlaw and Social Change: A Democratic Approach to Copyright Law in Cyberspace*, 14 CARDOZO ARTS & ENT. L.J. 215, 238 (1996) (referring to physical databases).

prioritization.[67] These choices make them sites of meaning-making, which impact collective memory. To use the words of Schwartz and Cook, "Through archives, the past is controlled."[68]

Often, these editorial choices are not easily transparent to the users exposed to the materials.[69] Yet, with respect to collective memory construction through algorithmic-based, seemingly human agents, this layer of editorial choices becomes even more invisible, due to the specific traits of the technology: the (almost) natural interaction, and the personal, intimate feelings that the medium evokes. In other words, the human tendency to anthropomorphize and attribute human qualities to social robots may further decrease our ability to discern between the real, and the virtual representation that is mediated through a series of human decisions.

A second prominent challenge that presents itself in the use of virtual memory agents and other AI-based technologies is their susceptibility to hacking, manipulation by third parties, or technical failures that may harm the authenticity of the mediated content. Obviously, similar vulnerabilities exist in any other field using AI and robots. Famous examples include Amazon's 'Alexa' proposing porn content to a toddler,[70] or reports about Microsoft's decision to suspend 'Tay,' its AI chatterbot, from tweeting, in light of its racist outbursts.[71] Moreover, the fear of manipulation subsists, to a certain extent, with respect to more traditional modes of memory mediation as well: documents can be forged, videos may be corrupted, and artifacts may be damaged. Yet, here, too, due to the traits of the medium, the concern becomes more pronounced, and the effect of misuse may be greater. Put differently, if our past is controlled through the archive,[72] our

67. *See, e.g.*, Hartman, *supra* note 22, at 201–02 (referring to the Yale testimony project and observing: "If we had stopped to resolve all the questions surrounding our effort—including that of the exact value of oral history as history—we would never have proceeded beyond the first experimental tapes. . . . We do not deceive ourselves into thinking that we have developed the perfect interview. There may be no such thing: the quality of oral history is influenced by the human chemistry between interviewer and interviewee, and even by the day and place of filming").

68. Joan M. Schwartz & Terry Cook, *Archives, Records, and Power: The Making of Modern Memory*, 2 ARCHIVAL SCI. 1, 1 (2002).

69. *Cf.* Shur-Ofry, *supra* note 62, at 322 (observing that the structures and organization choices of databases "diffuse into the information system" and users no longer pay attention to the "langue" of the database); GEOFFREY C. BOWKER & SUSAN LEIGH STAR, SORTING THINGS OUT: CLASSIFICATION AND ITS CONSEQUENCES 323 (1999) (referring to databases' structures and observing that for the user, the databases' structure often becomes "invisible").

70. *See* Judith Shulevitz, *Alexa, Should We Trust You?*, ATLANTIC (Nov. 2018), https://www.theatlantic.com/magazine/archive/2018/11/alexa-how-will-you-change-us/570844/ [https://perma.cc/KQ9R-KYWY].

71. *See* Kari Paul, *Microsoft Had to Suspend Its AI Chatbot After It Veered into White Supremacy*, VICE: MOTHERBOARD (Mar. 24, 2016, 11:21 AM), https://motherboard.vice.com/en_us/article/kb7zdw/microsoft-suspends-ai-chatbot-after-it-veers-into-white-supremacy-tay-and-you [https://perma.cc/7QAL-TZDH].

72. Schwartz & Cook, *supra* note 68, at 1.

future's past may be controlled through algorithms that mediate the contents of those archives. In such a (possible, feasible) future where most of our collective memory is mediated through AI and robots, manipulation can easily distort the memories of our joint past.[73]

The challenges we discuss above are salient, though certainly not exhaustive. Our analysis concentrates on the 'user side,' namely on the effect of AI-mediated technologies on users exposed to their output. Yet, AI mediation of collective memory raises additional ethical and legal questions, including questions pertaining to the relations between the AI creators and the 'contributors' of the underlying materials. When the database underlying the AI is created with the consent and active participation of those contributors, as was the case in the NDT project, these questions are not acute. Yet, they may arise if an AI based system uses pre-existing archives of photos, videos, or documents. Consider, again, the imaginary 'virtual Lincoln' that derives from a database of Lincoln's writings. A detailed discussion of this axis exceeds the scope of this Commentary, and certainly warrants future research.[74] Yet, the policy framework we propose in the following Part sheds light on these types of questions too.

What, then, is the appropriate policy response to the challenges that lie at the intersection of AI and collective memory? In the next Part, we introduce the concept of memory fiduciaries and argue that this conception allows us to devise an initial policy framework toward robotic collective memory.

III. TOWARD A FRAMEWORK OF MEMORY FIDUCIARIES

In a nutshell, we maintain that entities which employ robotic memory agents to mediate materials that affect our intergenerational memory should be considered 'memory fiduciaries,' and as such, should be subject to certain fiduciary duties.

What are the justifications for memory fiduciaries in the age of robotic collective memory? Who should be considered a memory fiduciary, and what type of duties should apply to these entities with respect to collective memory that is mediated by AI? The following Section presents the notion of memory fiduciaries and its justifications, while analyzing the parallels and differences between this concept and Balkin's information fiduciaries

73. *Cf.* Tal Z. Zarsky, *Privacy and Manipulation in the Digital Age*, 20 THEORETICAL INQUIRIES L. 157, 158 (2019) (discussing a risk of commercial manipulation entailed in digital environments).

74. *See, e,g,*, Pessach & Shur-Ofry, *supra* note 30 (analyzing copyright questions that arise when remembrance institutions seek to use pre-existing Holocaust-related materials). Notably, copyright is not the sole legal branch that may be relevant in such cases. Questions pertaining to privacy, the right of publicity, and additional legal doctrines may apply.

framework. The next Section then proceeds to demonstrate the application of the memory fiduciaries framework to the use of algorithmic memory agents. Our discussion, however, is far from comprehensive. While we do offer some tentative answers, our main purpose is to introduce memory fiduciaries as a potential policy framework for thinking about the challenges that lie at the intersection of collective memory and new technologies, and open up an avenue for future discussions of these issues.

A. From Information Fiduciaries to Memory Fiduciaries

Our proposed concept of memory fiduciaries is analogous, though not identical, to the concept of information fiduciaries. The latter was recently advanced by Jack Balkin as a response to rising privacy concerns in the digital age.[75] According to Balkin's analysis, the extensive collection and use of personal data in the digital age gives rise to new fiduciary duties that apply to digital organizations, even in the absence of a contractual basis.[76] These duties are based, to a large extent, on the *trust* that these entities "induce" from people, "[b]y presenting themselves as trustworthy collectors and keepers of our individual data."[77] The trust-based relationship is typically accompanied by *information asymmetries*, whereby the precise data collected and the ways in which it is used are not fully transparent to the ordinary user.[78] Therefore, entities collecting and using large amounts of personal data hold considerable *power*, which is only increasing as more and more decisions based on that data are made by algorithms.[79] This state

75. Balkin, *Information Fiduciaries*, *supra* note 10. For additional scholarship proposing to base privacy in the digital age on trust and fiduciary relationship, see Jessica Litman, *Information Privacy/Information Property*, 52 STAN. L. REV. 1283, 1308–09 (2000) (arguing that trust and fiduciary duties can serve as a basis for respecting privacy); Neil Richards & Woodrow Hartzog, *Taking Trust Seriously in Privacy Law*, 19 STAN. TECH. L. Rev. 431, 457–58 (2016) (arguing that "the concept of fiduciaries helpfully reorients privacy and crystalizes the concept of trust in information relationships"); ARI EZRA WALDMAN, PRIVACY AS TRUST: INFORMATION PRIVACY FOR AN INFORMATION AGE 85–92 (2018) (arguing that data collectors should be considered "information fiduciaries" due to the asymmetry and vulnerability that characterize their relationship with the users); DANIEL J. SOLOVE, THE DIGITAL PERSON: TECHNOLOGY AND PRIVACY IN THE INFORMATION AGE 103 (2004) ("[T]he law should hold that companies collecting and using our personal information stand in a fiduciary relationship with us."). This literature includes significant insights concerning the interface of privacy and trust, yet since our focus here is on intergenerational memory (and not on privacy), a comprehensive review of these discussions exceeds the scope of this Commentary.

76. Balkin, *Information Fiduciaries*, *supra* note 10*,* at 1200–05, 1221 (discussing the limits of the contractual model in response to privacy concerns in the digital age and proposing to overcome them via the concept of fiduciary duties that would apply even absent a contractual obligation).

77. *Id.* at 1223; *see also* Litman, *supra* note 75 (describing the trust-based relationship between entities that receive private information and the individuals that provide it); Richards & Hartzog, *supra* note 75 (promoting the concept of fiduciary duties, based on trust relationship, in engagements that involve the provision of personal information).

78. Balkin, *Information Fiduciaries*, *supra* note 10, at 1222–23.

79. *Id.* at 1185–86, 1232 ("As algorithms for making decisions based on this data become more powerful, so too will the people and organizations who collect and use the data.").

of affairs creates a reasonable expectation that the digital organizations collecting our data "will not betray us."[80] Rather, these entities have a duty to act in a trustworthy manner that is consistent with ethical standards and protect the trust we place in their hands.[81]

While Balkin's analysis concentrates on privacy concerns ensuing from the collection of individual data by private stakeholders, significant parallels can be drawn to the case of algorithmic collective memory. Most prominently, our relationship with the entities that mediate collective memory through algorithmic agents are also based on trust and confidence. We trust these entities that the materials they mediate through algorithmic agents are authentic materials, and that the deployment of such agents aims to recollect the past without external third-party interventions. To illustrate, let us briefly return to Pinchas Gutter and the NDT system. Witnesses like Gutter who contribute their testimonies as input for robotic memory agents trust that these testimonies will not be misrepresented but rather be used in a trustworthy manner. However, the trust relationship in this and similar cases is broader; it also extends to the side of the 'users.' Those exposed to robotic memory agents similarly trust that the output the virtual witnesses generate constitutes as accurate a representation as possible of the original testimonies.

Two important factors enhance trust in our case. The first is the human tendency to anthropomorphize. As the previous discussion indicates, when the past is conveyed through interactive social robots, which appear human and animated, feelings of empathy and trust are almost inevitable.[82] The second factor is more general and concerns our inclination to trust the entities that are active in the field of collective memory. Remembrance institutions, such as archives, libraries, museums, or memorials, collect and mediate to us much of the information that comprises collective memory.[83] These institutions are often not-for-profit entities that regard the collection and the provision of access to materials for purposes of intergenerational memory as their primary mission.[84]

---

80. *Id.* at 1222 ("[Online service providers] present themselves to the public as responsible and upstanding organizations who will use their power for lawful ends and, above all, who will not betray us.").

81. *Id.* at 1205–07, 1224–25. For a recent proposal of imposing specific duties on service providers who collect data on their users, see also Ariel Dobkin, *Information Fiduciaries in Practice: Data Privacy and User Expectations*, 33 BERKELEY TECH. L.J. 1 (2018).

82. *See supra* notes 53–56 and accompanying text.

83. *See* the discussion of remembrance institutions, *supra* notes 30–33 and accompanying text.

84. See, for example, the mission statement of the "American Memory" collection that forms part of the Library of Congress. *Mission and History*, LIBR. OF CONGRESS, https://memory.loc.gov/ammem/about/index.html [https://perma.cc/LR82-5EGE] ("American Memory provides free and open access through the Internet to written and spoken words, sound recordings, still and moving images, prints, maps, and sheet music that document the American experience. It is a digital record of American history and creativity. These materials, from the collections of the Library of Congress and other

This is not to say that business entities cannot be significant players in the field of collective memory. In fact, as technology develops, the impact of private technology companies on collective memory is bound to increase.[85] The Google Cultural Institute is a case in point: the Institute collaborates with more than a thousand museums and cultural institutions and, among other activities, presents digital exhibitions on topics central to collective memory, from the Fall of the Berlin Wall to the Apartheid.[86] However, even when market players are involved, these ventures are still perceived by the users, and often described by the providers, as public-oriented.[87] In other words, we have a social expectation that those who play in the field of collective memory adopt a public-oriented approach, and we may be more inclined to trust them *ex ante* in their activities in this area.[88]

Notably, a recent criticism of Balkin's information fiduciaries framework maintains that there could be a potential conflict between the market-oriented focus of technology companies and their duties toward their shareholders, and the proposed fiduciary scheme.[89] However, the long-term intergenerational nature of collective memory, and the public-oriented traits of its mediation, distinguish memory fiduciaries from technology platforms, and imply that those possible conflicts do not arise in an acute form in the present case. In other words, in the field of collective memory,

institutions, chronicle historical events, people, places, and ideas that continue to shape America, serving the public as a resource for education and lifelong learning."); *see also The History of the Shoah Memorial*, MÉMORIAL DE LA SHOAH, http://www.memorialdelashoah.org/en/the-memorial/presentation/the-history-of-the-shoah-memorial.html [https://perma.cc/M5K2-3SXW] ("Today the Memorial, a museum and archival center, is a place of mediation essential for the transmission of memory.").

85. Pessach, *supra* note 30, at 85 (arguing that the digital age brings with it increased privatization of memory institutions).

86. *See* GOOGLE CULTURAL INST., https://www.google.com/culturalinstitute/about/partners/ [https://perma.cc/7YBC-C6BT].

87. *See id.* ("Founded in 2011, the Google Cultural Institute is a not-for-profit initiative that partners with cultural organizations to bring the world's cultural heritage online."); *see also Google Books Library Project: An Enhanced Card Catalog of the World's Books*, GOOGLE BOOKS, https://www.google.com/googlebooks/library/ [https://perma.cc/9DTK-ZRVL] ("The Library Project's aim is simple: make it easier for people to find relevant books – specifically, books they wouldn't find any other way such as those that are out of print – while carefully respecting authors' and publishers' copyrights.").

88. *Cf.* JOSEPH L. SAX, PLAYING DARTS WITH A REMBRANDT: PUBLIC AND PRIVATE RIGHTS IN CULTURAL TREASURES (1999) [hereinafter SAX, PLAYING DARTS] (proposing that the administration of "cultural treasures" must consider the public interest, even when these treasures are placed in private hands); WHOSE MUSE? ART MUSEUMS AND THE PUBLIC TRUST (James Cuno ed., 2004) [hereinafter WHOSE MUSE?] (suggesting that museums must adopt a public-oriented approach); Guy Pessach, *The Role of Libraries in A2K: Taking Stock and Looking Ahead*, 2007 MICH. ST. L. REV. 257 (making similar observations with respect to libraries).

89. Lina M. Khan & David E. Pozen, *A Skeptical View of Information Fiduciaries*, 133 HARV. L. REV. (forthcoming 2019), https://ssrn.com/abstract=3341661. A thorough analysis of this argument is unnecessary in the present context, and exceeds the scope of this Commentary.

users' trust is a most salient feature, which further reinforces the adequacy of the memory fiduciaries framework.[90]

Moreover, power relations and information asymmetries that constitute important tenets in Balkin's concept of information fiduciaries prevail in our case too.[91] Whether public or private, remembrance institutions inevitably exercise discretion regarding which information to collect, record, preserve, digitize, and display, or conversely, leave out or marginalize. From the recipients' perspective, these decisions are often invisible and seldom questioned.[92] Yet, remembrance institutions "wield power over the shape and direction of . . . collective memory, and national identity, over how we know ourselves as individuals, groups, and societies."[93]

The unequal power relations and information asymmetries are likely to expand in the age of algorithms. One of the pillars of effective AI systems is the availability of large databases of relevant information.[94] In the case of collective memory, the relevant materials—documents, videos, or photos—are regularly held by remembrance institutions.[95] Ordinarily, individuals would not have the ability to access these raw materials in their entirety, nor the capacity to process them through AI-based technologies. Remembrance institutions may therefore be in the best position to develop algorithmic memory agents, possibly in cooperation with technology companies.

To a large extent, then, one can expect that those who "control the archives" will also control the algorithms that mediate their contents to the public. It is entirely plausible that the algorithms themselves will not be publicly shared. Even in cases like the NDT project, whose creators demonstrate considerable openness regarding the design principles underlying the virtual witnesses system, some parts of the system—for example the natural language processing software or the stock of underlying answers—are, to the best of our knowledge, not open to the public. And even if they were, it is unrealistic that ordinary members of the public would possess the capacity or will to scrutinize them.[96]

---

90. *Cf. id.* (manuscript at 33) (arguing that the salient feature of companies like Google and Facebook is "not that their end users must be able to trust and depend on them").

91. For the discussion of these factors under the information fiduciaries framework, see *supra* notes 77–80 and accompanying text.

92. *See supra* notes 62–69 and accompanying text; Schwartz & Cook, *supra* note 68, at 3.

93. Schwartz & Cook, *supra* note 68, at 2.

94. *Cf.* Press Release, European Comm'n, Artificial Intelligence: Commission Outlines a European Approach to Boost Investment and Set Ethical Guidelines (Apr. 25, 2018), http://europa.eu/rapid/press-release_IP-18-3362_en.htm [https://perma.cc/5SJ8-M926] ("[D]ata is the raw material for most AI technologies . . . .").

95. *See supra* notes 30–33 and accompanying text; *cf.* Pessach & Shur-Ofry, *supra* note 30, at 136, 168–69 (explaining that market stakeholders would seldom be interested in such materials).

96. This phenomenon is of course not unique to robotic memory agents. Similar asymmetries exist in additional fields that involve algorithmic-based decisions. *See, e.g.*, FRANK PASQUALE, THE

Beyond those common features of trust, power relations, and asymmetries, collective memory and information share two fundamental traits that further justify our proposal. First, similar to information, collective memory is a *public good*, with qualities of non-rivalry and non-excludability that make it vulnerable to misuse and manipulation.[97] As the preceding analysis instructs, these fears are intensified in the age of robotic collective memory.[98] Therefore, classical public-goods theory reinforces the need for policy interventions designed to prevent such misuse.[99] Secondly, our discussion of collective memory illuminates its substantial social value and its importance for the formation of social and individual identities.[100] In light of these traits, collective memory forms a vital part of cultural democracy. Just as the participation in contemporary discourse is essential for shaping people's identities, worldviews, and beliefs, so is the participation in the intergenerational discourse. To use the renowned words of Jacques Derrida, "[t]here is no political power without control of the archive, if not of memory."[101]

Balkin's submission that "[p]eople have a right to participate in forms of power that reshape and alter them because what is literally at stake is their own selves,"[102] is equally convincing with respect to collective memory, particularly when it is mediated through AI and algorithms. These insights imply that users' participation in the collective-memory discourse is protected by their freedom of expression, and provide further justification for imposing certain duties on the entities that possess the power to mediate

---

BLACK BOX SOCIETY (2015) (discussing the asymmetries between the ample data which corporations collect on users while employing powerful algorithms, and the obscurity regarding the use of that data by the corporations themselves); Balkin, *Three Laws*, *supra* note 52, at 1227 (arguing that the use of algorithms should be subject to obligations of transparency, due process, and accountability); Joshua A. Kroll et. al, *Accountable Algorithms*, 165 U. PA. L. REV. 633 (2017) (introducing computer science concepts that can be used to set out and verify algorithmic compliance with standards of legal fairness for automated decisions); Deven R. Desai & Joshua A. Kroll, *Trust But Verify: A Guide to Algorithms and the Law*, 31 HARV. J.L. & TECH. 1, 3–4 (2017) (discussing various concerns that arise due to the "black box" nature of algorithmic decisions).

97. *Cf.* Pessach, *supra* note 30, at 116 ("[M]emory institutions and their subject matters are regarded by many as *public goods* . . . ."); Richard S. Whitt, *"Through a Glass, Darkly": Technical, Policy, and Financial Actions to Avert the Coming Digital Dark Ages*, 33 SANTA CLARA COMPUTER & HIGH TECH. L.J. 117, 178–79 (2016) (identifying cultural preservation as a public good). For a general discussion of the nature of public goods, including their non-rivalrous and non-excludable qualities, see, for example, RICHARD CORNES & TODD SANDLER, THE THEORY OF EXTERNALITIES, PUBLIC GOODS, AND CLUB GOODS 3–4 (1986); Joseph E. Stiglitz, *Knowledge as a Global Public Good*, *in* GLOBAL PUBLIC GOODS: INTERNATIONAL COOPERATION IN THE 21ST CENTURY 308, 308–10 (Inge Kaul, Isabelle Grunberg & Marc A. Stern eds., 1999).

98. *See supra* note 73 and accompanying text.

99. *See, e.g.*, Stiglitz, *supra* note 97, at 311 (explaining that the protection and provision of public goods require certain state intervention).

100. For the social value of collective memory, see *supra* notes 17–19 and accompanying text.

101. JACQUES DERRIDA, ARCHIVE FEVER: A FREUDIAN IMPRESSION 4 n.1 (Eric Prenowitz trans., 1996).

102. Balkin, *Information Fiduciaries*, *supra* note 10*,* at 1211.

collective memory, under the memory fiduciaries framework we advocate here.[103] They also provide the answer to First Amendment concerns that were raised in the context of information fiduciaries, and clarify that placing reasonable duties on remembrance institutions in order to ensure people's rights to participate in the construction of collective memory does not contradict the rights of those institutions under the First Amendment.[104]

Altogether, the foregoing analysis indicates that there are solid justifications for extending the information-fiduciaries approach to the realm of collective memory, and specifically to the use of algorithmic memory agents. We therefore propose to recognize that certain fiduciary duties apply to the collection, provision, and mediation of materials that affect our intergenerational memory, through algorithmic memory agents. We refer to the entities that are subject to these duties as "*memory fiduciaries*."

Before we proceed to discuss the nature of memory fiduciaries and describe their duties in the case of algorithmic collective memory, a clarification is in order. Despite the substantial parallels, the concept of information fiduciaries is not completely identical to memory fiduciaries. As explained, Balkin's framework was developed against a reality of powerful digital organizations—Facebook, Uber, or Google, to name some of his examples—that collect and control individuals' data. The beneficiaries of the fiduciary duties under his theory are therefore the individuals whose information is being collected and processed.[105] In the case of memory fiduciaries, the picture is more complex. Remembrance institutions that collect information and materials, which then become building blocks in the construction of collective memory, certainly owe fiduciary duties to individuals, like Pinchas Gutter, who contribute these materials. Yet, under our proposed framework, their duties extend beyond the contributors, and apply, in addition, to the *recipients* of their output—in our example, to users exposed to the virtual Gutter testimony.

Oft-times, the relationship between these recipients, or users, and the providers of the information will not be an explicit contractual relationship. Moreover, from the institutions' perspective, these users are not necessarily identified individuals. Nevertheless, our analysis demonstrates that trust,

---

103. *Cf.* Jack M. Balkin, *Cultural Democracy and the First Amendment*, 110 NW. U. L. REV. 1053 (2016) (explaining that under a cultural-democracy theory, the rights of people to participate in the formation of culture and meaning-making processes are protected as part of freedom of expression).

104. *Cf.* Balkin, *Information Fiduciaries*, *supra* note 10, at 1225 ("[R]easonable obligations placed on information fiduciaries do not violate the First Amendment . . . ."); Pasquale & Cockfield, *supra* note 7, at 864–66 (noting that "bot-mediated communication is an entirely distinct phenomenon from previous modes of communication" and warning against privileging algorithmic data-processing as speech).

105. Balkin, *Information Fiduciaries*, *supra* note 10, at 1233–34.

power, and asymmetries certainly characterize the relationship between those who mediate collective memory and the recipients of their outputs. Furthermore, due to the technological traits of algorithmic memory agents, this relationship is particularly vulnerable. Therefore, while the concept of memory fiduciaries may require a certain extension of extant doctrine, we believe this extension is well justified, and that the doctrine of fiduciary duties can and should be developed to accommodate memory fiduciaries.[106]

We certainly acknowledge that the fiduciary framework we explore here is not the sole legal construct that can apply to the field of collective memory. Additional legal doctrines, such as public trust, can be useful in some cases, especially when the relevant entities are public entities and their activities affect the public at large.[107] Yet, given the attributes of the relationship we describe in the foregoing paragraphs, the diverse nature of the entities that may employ algorithmic memory agents, and the inherent flexibilities of the fiduciary doctrine itself, we believe this doctrine provides a natural starting point for devising an appropriate governance framework for robotic collective memory. Our approach, which relies on the common-law concept of fiduciary, is also consistent with recent scholarly approaches that illuminate the advantages of flexible common-law frameworks for regulating challenges pertaining to AI and robots.[108]

The next Section proceeds to explore the duties that apply to memory fiduciaries in relation to robotic collective memory, while highlighting a series of related questions.

---

106. For the flexibility of the doctrine, see, for example, Tamar Frankel, *Fiduciary Law in the Twenty-First Century*, Essay, 91 B.U. L. REV. 1289, 1299 (2011) (acknowledging that fiduciary law can be a useful model in various contexts); TAMAR FRANKEL, FIDUCIARY LAW 9 (2011) (explaining that the degree of entrusted power determines whether a fiduciary relationship exists and how strict the duties should be); Balkin, *Information Fiduciaries*, *supra* note 10, at 1223 ("[A] changing society generates new kinds of fiduciary relations and fiduciary obligations that the law can and should recognize.").

107. For the classical doctrine of 'public trust,' see, generally, Joseph L. Sax, *The Public Trust Doctrine in Natural Resource Law: Effective Judicial Intervention*, 68 MICH. L. REV. 471 (1970) (explaining that under the principle of public trust, the State acts as a trustee for the public with respect to certain common resources, such as the seashore, highways, and running water). For proposals to apply the public trust principles to the cultural sphere, see, for example, SAX, PLAYING DARTS, *supra* note 88 (discussing the social responsibilities of museums and their long-standing function as public trusts); *cf.* John Nivala, Droit Patrimoine*: The Barnes Collection, The Public Interest, and Protecting Our Cultural Inheritance*, 55 RUTGERS L. REV 477, 541–44 (2003) (discussing the case of the Barnes collection and the justifications to impose responsibilities and obligations on museums, e.g. regarding the public's access to their collections, based on their public functions). A full exploration of the applicability of these, and additional legal doctrines, to our case warrants further research and exceeds the scope of the present discussion.

108. *See* Lemley & Casey, *supra* note 51 (manuscript at 6) ("[A] common law, case-by-case approach may provide a promising means of successfully navigating the definitional issues presented by robots—one that builds and adapts its definitions inductively over time rather than trying to legislate it.").

B. Memory Fiduciaries in the Age of AI

Preliminary clarifications are in order regarding the contours of our discussion and its location in the broader landscape. This Commentary focuses on the intersection of collective memory and AI. The analysis in the previous Parts instructs that this interface raises new and pressing challenges, which make the fiduciary framework particularly apt. The following paragraphs apply our proposed framework to memory fiduciaries that mediate the past through algorithmic memory agents, and sketch, in broad strokes, their fiduciary duties. However, the memory-fiduciaries framework could possibly have more general implications for collective memory, beyond its interface with robots, AI, or other cutting-edge technologies. The nature of collective memory as a public good that is susceptible to vulnerabilities, and as an essential part of cultural democracy, implies that this conceptualization may be useful in analyzing broader questions pertaining to collective memory construction through more traditional means. In other words, the memory fiduciary concept can possibly inform the analysis of diverse questions related to intergenerational collective memory, from access to historical materials through preservation of physical "memory sites," to the display of historically significant artifacts. However, exploring the potential of the memory-fiduciaries framework to address such broader issues exceeds the scope of this Commentary. We leave it for future work.

We should also clarify that the memory-fiduciaries framework is not confined to public or quasi-public remembrance institutions. Of course, remembrance institutions, such as museums, archives, and libraries, with substantial activities of documenting, preserving, and providing access to various materials should be subject to this framework. However, the increasing involvement of private business players and technology companies in the field of memory construction prescribes that these entities, too, could be regarded as memory fiduciaries with regard to some of their activities—consider, again, the example of the Google Cultural Institute.[109] It is reasonable to assume that these entities may already view themselves as subject to some sort of duties in their activities in the field of collective memory, even if they do not use the term fiduciary to describe them.[110] Indeed, border cases may arise, and line-drawing would be inevitable under our framework, as it is under any other legal construct. Yet, as a general

109. *See supra* notes 86–87 and accompanying text.

110. *See* the discussion of the public-regarding approach that remembrance institutions often adopt, *supra* notes 83–87 and accompanying text; *cf.* Sundar Pichai, *AI at Google: Our Principles*, GOOGLE (June 7, 2018), https://www.blog.google/technology/ai/ai-principles/ [https://perma.cc/CG8G-QABW] (Google's AI principles include, *inter alia*, principles relating to security and accountability, including the provision of "explanations").

guideline, the application of the memory fiduciaries framework would not depend on the public or private nature of the entity, but on the activities of that entity in the field of collective memory and the attributes of its *relationships* with both contributors and end-users.[111]

Which duties, then, should apply to memory fiduciaries that mediate collective memory through algorithmic agents? The notion of fiduciary generally consists of two main obligations: a duty of loyalty—also described as "trustworthiness" or a duty "not to betray"—and a duty of care, or "a duty not to harm."[112] The concrete contents of these duties depend on the precise circumstances. The circumstances in the case of deploying robotic agents for mediating collective memory illuminate two principal challenges: the subsistence of unavoidable editorial choices that are invisible to the end-users, and the susceptibility of the technology to misuse, hacking, or manipulation by third parties.[113] The design of memory fiduciary duties should therefore correspond to these challenges.

Concomitantly, we must also take into account the benefits entailed in the use of algorithmic memory agents, and shape fiduciary duties in a way that will not jeopardize their deployment.[114] Bearing in mind these general contours, we suggest four prominent, though not exhaustive, duties that apply in the case of robotic collective memory: 1) 'Identify the Robot'; 2) Explainability; 3) Reasonable Data Security; and 4) Integrity.[115]

Under the *first principle*, entities that mediate collective memory through AI should be under an obligation to inform end-users exposed to the outputs that AI is involved.[116] The teenagers who interact with the virtual Gutter in a classroom surely understand that they are sitting in front of a screen. Yet,

---

111. *Cf.* Balkin, *Information Fiduciaries*, *supra* note 10, at 1186–87 (suggesting that the duties of digital organizations collecting information about users should derive from the nature of the relationship). Notably, this Commentary does not purport to strictly draw the definitive borders of memory fiduciaries. Hopefully, future work will further develop the distinctions we propose here. However, as we explain below, the common-law nature and case-sensitivity of the fiduciary doctrine will allow us to resolve border-line cases as they arise.

112. FRANKEL, *supra* note 106, at 106–07, 169 (further suggesting that the duty of care is weaker than the duty of loyalty); *see also* Balkin, *Information Fiduciaries*, *supra* note 10, at 1222 (describing information fiduciaries' duty "not to betray").

113. *See supra* notes 61–74 and accompanying text.

114. For a discussion of the benefits, see *supra* notes 53–60 and accompanying text.

115. Notably, we do not argue that the interface of AI and memory necessarily requires devising unique solutions *de novo*, and indeed some of the principles we propose here are derived from the vast literature that investigated other challenges posed by AI and robots and offered a broad menu of regulatory responses. *See, e.g.*, *Universal Guidelines for Artificial Intelligence*, PUB. VOICE (Oct. 23, 2018), https://thepublicvoice.org/ai-universal-guidelines [https://perma.cc/B3JY-MDRV]. Rather, our aim is to sketch the contents of the proposed fiduciary duty, by highlighting specific principles that correspond to the particular challenges that lie at the intersection of AI, robots, and collective memory.

116. *See* Tim Wu, *Please Prove You're Not a Robot*, N.Y. TIMES (July 15, 2017), https://www.nytimes.com/2017/07/15/opinion/sunday/please-prove-youre-not-a-robot.html [https://perma.cc/ZFT3-SZKW] (proposing a "Blade Runner" law that prohibits the use of "any program that hides its real identity to pose as a human").

those watching a video documentation of the interaction may not be so certain as to the virtual nature of the witness.[117] And as visualization technologies develop, our ability to distinguish between the real and the virtual will likely further diminish.[118] Notably, AI employed in mediating collective memory may not always have a visual interface—imagine, for example, a conversation with the 'ultimate witness' of the Vietnam War, an AI, voice-based, interactive system that answers questions about the War, using an integrated database of hundreds of testimonies. In such cases, it may be nearly impossible to identify the robotic nature of the system absent an explicit notification.[119]

In broader terms, the more algorithmic collective memory becomes ubiquitous and widely distributed through myriad digital channels, the greater the need is to alert users of its robotic nature. We should perhaps clarify that the principle of 'identify the robot' is not equivalent to 'define the robot.' Literature instructs that such definition may be difficult, at times even impossible, and it is not required for our purposes.[120] Rather, our emphasis is essentially on the need to exercise transparency when collective memory is mediated through AI-based technologies.

Under the *second*, related principle, a duty of explainability should apply to entities using AI to mediate collective memory. The notion of 'explainability,' derived from administrative law, has recently penetrated the field of AI. According to Frank Pasquale, explainability in this context suggests the provision of "a clear sense of the history of a robot—how was it first programmed, to what has it been exposed, and how has this interplay between hardware, software, and the external environment resulted in present behavior."[121] In our case, explainability similarly implies a duty of entities that use algorithmic memory agents to provide information about the system's general design principles. For example, what is the database of raw materials that were used to create the system, and what is the relationship between the underlying materials and the output generated by the virtual memory agent? We do *not* propose, however, that explainability includes a general duty to provide the code underlying the system. Such an obligation would, in most cases, be cumbersome for the memory fiduciary

117. Consider the video of the discussion quoted in the Introduction *supra* note 1.

118. *See* the examples *supra* notes 43–47 and accompanying text.

119. Consider Google's Duplex technology *supra* notes 43–44 and accompanying text.

120. For the definitional challenges, see Lemley & Casey, *supra* note 51, and the literature cited *supra* notes 51–52 and accompanying text.

121. *See* Frank Pasquale, *Toward a Fourth Law of Robotics: Preserving Attribution, Responsibility, and Explainability in an Algorithmic Society*, 78 OHIO ST. L.J. 1243, 1252 (2017).

and have little value for the end user.[122] Rather, it should focus on explaining the *principles* underlying the design.

To illustrate how explainability might work in the field of algorithmic collective memory, the NDT project is again a case in point. The USC Shoah Foundation released information that describes how the database of questions and answers underlying the virtual agents was created, including the general number of questions the survivors were asked, as well as the location, setting, and duration of the filming.[123] It clarified that the interface of the system and the end-user is controlled by a natural language processing software. It further explained the relations between the answers of the virtual survivors and the underlying database. According to the information provided, the system draws the virtual survivor's answers from the database of original answers that the survivor provided, and plays back these answers verbatim to the end user.[124]

Future developments may complicate the picture and further reinforce the need for explanation. As the technology of algorithmic memory agents becomes prevalent, it is far from certain that other remembrance institutions will adopt NDT's expansive approach towards information provision. In addition, with the rapid technological advancement of AI, the use of algorithmic memory agents could develop in additional directions. For example, researchers anticipate a significant qualitative improvement in the ability of machines to 'understand' information rather than just index it. The expectation is that in future years, AI that is able to synthesize information, 'connect the dots,' and generate evaluations and knowledge about the world will become an integral part of various social activities.[125] Consider, then, a virtual memory agent whose output does not merely consist of an accurate, verbatim representation of underlying materials, but who can "fill in" gaps in those materials, and generate new answers that are based on synthesizing, summarizing, and "understanding" them.

To illustrate, let us return to the hypothetical 'ultimate witness,' a voice-based AI system that delivers an integrated testimony about the Vietnam War based on myriad recorded and written testimonies.[126] Imagine, that in

122. *Cf.* Kroll et al., *supra* note 96, at 639 (noting that full transparency and exposure of software code "may still fail to resolve the concerns of many participants").

123. *See supra* notes 36–42 and accompanying text.

124. *Id.*

125. *See, e.g.*, Stuart Russell, *Q&A: The Future of Artificial Intelligence*, http://people.eecs.berkeley.edu/~russell/temp/q-and-a.html [https://perma.cc/KR6Z-4XVC] ("[A]s machines improve their grasp of language, search engines and 'personal assistants' on mobile phones will change from *indexing* web pages to *understanding* web pages, leading to qualitative improvements in their ability to answer questions, synthesize new information, offer advice, and connect the dots. . . . [S]ystems that know and reason about the real world, not just repositories of data—will become integral parts of society.").

126. *See supra* note 65 and accompanying text.

answering users' questions the virtual witness does not merely quote the relevant passages from existing testimonies, but rather synthesizes the sources to phrase new, more comprehensive answers, and even to raise hypotheses that fill in gaps in the raw materials. Obviously, in such cases the element of editorial discretion would be much more significant in comparison to the virtual Gutter, and the discrepancy between the raw materials and the output substantially larger. As a minimum, the duty of explainability would mandate the disclosure of the design principles underlying such virtual agents, including the existence of a machine-learning element that affects its output. More broadly, this example demonstrates that the exact contents and breadth of the explainability duty, and of memory fiduciaries' duties more generally, could vary under different circumstances. While we cannot sketch here every potential scenario, as a general matter, where the perceived authenticity of the memory-mediation is greater these duties should be heavier.

A duty of explainability may also nudge remembrance institutions contemplating the introduction of algorithmic memory agents to engage in critical self-evaluation of the technologies they wish to employ and the entailed risks and benefits, without interfering with their editorial discretion. As such, explainability could create a *de facto* standard of conduct and serve not only as a governing principle but also as a normative source for remembrance institutions deliberating the increasingly difficult dilemma at the interface of AI and collective memory.

From the users' perspective, recognizing a duty to 'identify the robot' together with a duty of explainability will raise users' awareness that their interactions entail a mediated representation of the original testimony or other raw materials, protect them from mistakenly assuming that they are interacting with a human, and shield them from feeling betrayed.[127] It will also direct a spotlight to the otherwise invisible layers of editorial choices embedded in the deployment of algorithmic agents, and allow users to 'calibrate' their judgement and understanding accordingly. Clarifying that memory is mediated through AI is therefore consistent with the principles of loyalty and care that constitute the core of fiduciary duties. Concomitantly, research instructs that the tendency to anthropomorphize social robots does not disappear when people are aware that they are interacting with a robot, even when they understand the robot's underlying mechanism.[128] Therefore, abiding by these principles will not jeopardize the

127. *Cf.* Sherry Turkle, *Authenticity in the Age of Digital Companions*, *in* MACHINE ETHICS 62 (Michael Anderson & Susan Leigh Anderson eds., 2011) (arguing that authenticity and the difference between social robots and sentient creatures are still, and should be, important to people).

128. *See supra* note 56 and accompanying text, which indicates that anthropomorphism subsists even among professional computer scientists.

users' experience of personal interaction, nor undermine feelings of empathy.

Our preceding analysis further instructs that algorithmic memory agents are vulnerable to risks of hacking, or third-party manipulation that can harm the integrity of the mediated content. Therefore, under our proposed *third principle*, memory fiduciaries should introduce reasonable data security measures to prevent those risks.[129] To illustrate, let us return once more to the virtual Pinchas Gutter. When confronted by a reporter with a direct question concerning Holocaust denial, the virtual Gutter answered: "To someone who has spent five years in hell—a living witness—[people who say] that this did not happen . . . I believe that they are just as bad as the perpetrators . . . . Every one of them should be taken to a court of law because they are in contempt of humanity itself."[130] It is easy to hypothesize how a hacked version of the system could provide a very different answer, which in turn could have a distorting effect on collective memory. Put differently, the use of algorithmic memory agents creates an enhanced risk of manipulation, and memory fiduciaries who decide to employ these agents are under a duty to adopt protection measures to internalize this risk. This duty, too, is consistent with principles of harm prevention and loyalty, not only toward the end-users but also to contributors like Gutter, who trust memory fiduciaries to preserve the integrity of their testimonies.

This last example leads us to a related, *fourth principle*, which we call *integrity*. It could be insufficient to identify the robot, explain the principles underlying the system, and maintain its data security, if the mediated contents are intentionally distorted—to illustrate, consider a hypothetical example of a museum that uses virtual witnesses to promote a Holocaust-denial narrative. The principle of integrity implies that when memory fiduciaries mediate collective memory through robots and AI, they need to apply these automation processes in a way that conforms with standards of accuracy and aspires to be consistent with historical facts.

Indeed, as our preceding discussion instructs, collective memory, by its nature, lends itself to a multiplicity of voices and does not, in fact cannot, constitute an objective historical account. Rather, each process of mediating

---

129. For somewhat similar proposals in the context of access to personal data, see, for example, Pasquale & Cockfield, *supra* note 7, at 863 (discussing the duty to introduce legal protection against "outside hackers and others who are interested in illegal or improper access to personal data"); *see also Universal Guidelines for Artificial Intelligence*, *supra* note 115 ("Institutions must secure AI systems against cybersecurity threats.").

130. *See* Emanuel Maiberg, *In the Future, the Holocaust Is Just Another Hologram*, VICE: MOTHERBOARD (Apr. 25, 2017, 9:44 AM), https://www.vice.com/en_us/article/ez3m4p/in-the-future-the-holocaust-is-just-another-hologram [https://perma.cc/R8QV-ZN57] (first and second alteration in original) (the quote refers to the answer given by the virtual witness in response to the reporter's provocative question: "I don't believe the gas chambers existed").

collective memory inevitably involves editorial choices and subjective discretion.[131] Nevertheless, while memory fiduciaries certainly face a broad continuum of legitimate editorial choices, they are under a duty not to distort the mediated materials, nor to knowingly deviate from accepted historical truths. Clearly, the principle of integrity is not limited to collective memory mediated by robots, and applies to remembrance institutions that mediate collective memory through more traditional media as well. However, as our discussion above demonstrates, the reliance of users on the robotic outputs is especially pronounced, and therefore the corresponding duty of integrity is particularly significant in this context.

Taken together, the four principles we explored will allow maintaining the benefits of robotic memory agents while minimizing the costs involved. These principles, however, are not exhaustive. We do not purport to address all possible scenarios that may arise as algorithmic memory agents become widespread, nor do we argue that the fiduciary framework provides a magical solution to all the challenges pertaining to AI and collective memory. Our aim is to illustrate how the memory fiduciaries notion provides a helpful framework for thinking about some of these challenges and can be applied to handle them. The inherent flexibility of the fiduciary doctrine and its sensitivity to context would allow courts and scholars to develop it on a case-by-case basis, to accommodate additional duties, or reshape these duties to address the myriad additional questions that are bound to arise at the intersection of AI and collective memory.

CONCLUSION

It is by now conventional wisdom that AI will deeply affect our future. This Commentary suggests that its impact on our collective past may be no less significant. Our analysis demonstrates that the intersection of AI, robots, and collective memory carries enormous potential, alongside substantial challenges. These challenges have been largely overlooked by the law and policy discussions of AI, social robots, and robot-human interface. Nevertheless, they deserve consideration by policy makers.

Our analysis further indicates that confronting these challenges and designing appropriate policies for the era of algorithmic collective memory must derive from, and connect, two disparate strands of knowledge. On the one hand, it warrants a close and detailed look at the technologies underlying robotic memory agents. On the other hand, it requires an exploration of collective memory, its traits and significance for the formation of social and individual identities.

131. *See supra* notes 20–21, 62–69 and accompanying text.

This Commentary has taken the first steps in combining these strands, and introduced the concept of memory fiduciaries as a policy framework for addressing robotic collective memory. Hopefully, future work will further develop and calibrate this framework. More broadly, collective memory and new technologies form a curious site of intersection of past and future. The legal regulation of this site deserves further exploration. This Commentary, we hope, will spark a conversation in this direction.