Selected Papers of #AoIR2020:
The 21st Annual Conference of the
Association of Internet Researchers
Virtual Event / 27-31 October 2020

HUMAN-MACHINE WRITING AND THE ETHICS OF LANGUAGE MODELS

Heidi A. McKee
Miami University

James E. Porter
Miami University

Our lives are intimately connected with networked writing machines. We write to, for, and with machines—and machines write to, for, and with us. Increasingly, much of what we read online is written by machines, such as the newswriting bot Heliograf, Persado's marketing copy app, Narrative Science's GameChanger app, and chat and twitterbots too numerous to name. And our emails and other communications are often co-written with machine writing agents, such as Google's Smart Compose, which suggests words, phrases and sentences for us to use.

The research we summarize here focuses on the language models or "informational frameworks" (Russo, 2018) that support machine writing agents, especially AI-based writing agents. How do these systems accumulate data?, how do they learn?, what are their principles and procedures for generating text?, is there anything missing from the model necessary for effective and ethical communication?, and, importantly, what ethical codes do these language models embody? Our ethics framework draws from the field of machine ethics (e.g., Dignum, 2018; Floridi, 2018; Leikas et al., 2019; Malle, 2016), as well as communication/language theory (e.g., Bengtsson, 2018; Chai et al., 2016; McKee & Porter, 2017; McKee & Porter, 2020).

Ethical Assumptions in Language Models

How AI agents operate is, of course, shaped by how they learn language in the first place. Often, AI writing systems employ a language model based on a reductive,

Suggested Citation (APA): McKee, Heidi A., & Porter, James E. (2020, October). *Human-machine writing and the ethics of language models*. Paper presented at AoIR 2020: The 21th Annual Conference of the Association of Internet Researchers. Virtual Event: AoIR. Retrieved from http://spir.aoir.org.

formalist model of text generation. For example, the Talk to Transformer app takes a short textual prompt and, using GPT-2, an OpenAI generative language model, creates a written article from that prompt, using a predictive model that creates new text based on the preceding text (OpenAI, 2019; Radford et al., 2019). This linear model (aligned with the Shannon-Weaver model) is based on a questionable core assumption: that meaning arises solely based on combining topic knowledge (derived from a database) with a grammatical/syntactical notion of text coherence.

But what is missing in this model is the crucial element shaping any communication: the context, including audience, exigence, purpose, and ethical understandings. The linear model's approach to coherence is based on a formalist notion of textual coherence (does one piece of text follow topically from another?) versus a social notion of coherence that involves human participation (does the text make sense to some intended readers in the contexts of interaction? does the text serve a meaningful social purpose?). The ethical implication of this linear model is that communication is a one-way process in which the communicator is the expert who "packages" knowledge or information into a text, which is then transmitted to the uninformed audience, who contributes little or nothing to knowledge formation.

The linear model contrasts with a more social/participatory ethical model, which begins by acknowledging the vital contribution of the audience to meaning making and, further, by recognizing that the exigence for communication arises from audience in the first place. Communication and meaning-making are collaborative processes. An ethical approach recognizes that meaning-making is constructed by interlocutors within the rhetorical context in which their communications circulate (McKee & Porter, 2020).

Examples of Machine Writing Agents

Writing machines will never write ethically if they are simply writing for textual coherence, An extreme example of this problem is Microsoft's Tay, who was grammatically and syntactically correct but who also, in less than 24 hours, became a racist, xenophobic, anti-Semitic, anti-feminist supporter of White Nationalism because the machine had not been designed or programmed with context in mind.

But not all machine writers fail. They are generally effective handling very well-defined tasks with established genre conventions, clearly identified audience needs, and predictable interaction scripts (e.g., customer service bots who offer set solutions to common problems). Not surprisingly, where machine writers struggle most is in open-ended situations with multiple audiences, competing needs, unclear expectations, complex contexts, and/or incomplete data.

A well-known AI agent is Narrative Science's Quill, which powers the GameChanger app that writes recaps for youth sports, taking the game box score and converting it into a narrative story. The linear model shaping GameChanger—if X in box score, write Y—limits its reporting, in ways that if you were at the game, can be quite humorous. The box score reads Derek Johnson 1 HR is translated into "Derek J. smashed a home run," converting the data into a full sentence with an action verb. But that home run was no

"smash," it was a dribbler that the infielders and outfielders muffed. Without the full context, it's hard for GameChanger to actually tell an accurate story of the game. Youth sports games are not high stakes events, but what happens when AI is the sole reporter for more significant events—elections, for example, or protest marches? The limitations regarding context become much more significant and ethically problematic as communication stakes are higher.

Bottom line is that context matters and not all AI systems can navigate context adequately. In our interview with Dennis Mortensen, CEO and founder of x.ai and creator of Amy Ingram, the online, AI-based scheduling assistant, he discussed the complexity of context in something as seemingly simple as setting up a meeting: "Even when [humans] speak to time, sometimes it doesn't even look like time. They'll say things like, 'Let's meet up later.' Later? What does that mean?" Preparing AI agent Amy for market took years and required over 65 human trainers because, as Mortensen explained to us, "We need[ed] to train on ambiguity. Amy needs to exist in your [human] universe." What Mortensen realized was that his AI-writing system needed human help—it needed human common sense (Hao, 2020) and practical reason (Marcus & Davis, 2020), it needed to understand the exigencies of the situation and the needs of human participants, and it needed the capacity for human inference-drawing (i.e., the ability to determine what is needed even if it is not clearly and explicitly articulated).

There are some AI-based writing systems that assess audience needs. Figaredo (2020) discusses how an AI-based learning management system can collect information abouts its student audience using the LMS itself as a database informing its choices about "how to adapt course design to student needs." Similarly, the AI-based marketing app Persado collects information about users to predict what kind of marketing messaging will be effective with different customers. It, too, has some contextual awareness of audience built into its production system. We say "some awareness," because it would be dangerous assume that the data collected adequately captures readers' needs or that the data is implemented ethically. Merely drawing from "big data" does not guarantee that the data are relevant or right or used with the audience's benefit in mind: big is not necessarily helpful or ethical; it can be manipulative.

Conclusions

(1) For now machine writing systems still need humans—particularly for the ethical guidance required in any communication interaction.

(2) Human writers need to understand the affordances and limitations of their machine writing assistants. Where and how can the AI writing systems be useful—and maybe at times even better than the human writer? Where and when do they need human writer intervention?

(3) AI-writing system designers need to critically examine the language models they are building for these systems. We need to move beyond formalist, linear input models to more complex social and contextual models that account for the broader and, yes, messier contexts in which communication arises and circulates.

References

Bengtsson, Stina. (2018). Ethics exists in communication: Human-machine ethics beyond the actor-network. Media@LSE Working Paper Series, London School of Economics and Political Science. OAI: DiVA.org:sh-37239

Chai, Joyce Y., Fang, Rui, Liu, Changsong, & She, Lanbo. (2016). Collaborative language grounding toward situated human-robot dialogue. *AI Magazine*, 37(4), 32–45. DOI: https://doi.org/10.1609/aimag.v37i4.2684

Dignum, Virginia. (2018). Ethics in artificial intelligence: Introduction to the special issue. *Ethics and Information Technology*, *20*(1), 1-3. DOI: https://doi.org/10.1007/s10676-018-9450-z

Figaredo, Domínguez D. (2020). Data-driven educational algorithms pedagogical framing. *RIED. Revista Iberoamericana de Educación a Distancia*, *23*(2), 65-84. doi: http://dx.doi.org/10.5944/ried.23.2.26470

Floridi, Luciano. (2018). Semantic capital: Its nature, value, and curation. *Philosophy & Technology*, *31*, 481-497. DOI: https://doi.org/10.1007/s13347-018-0335-1

Hao, Karen. (2020, January 31). AI still doesn't have the common sense to understand human language. *MIT Technology Review*. https://www.technologyreview.com/s/615126/ai-common-sense-reads-human-language-ai2/

Leikas, Jaana, Koivisto, Raija, & Gotcheva, Nadezhda. (2019). Ethical framework for designing autonomous intelligent systems. *Journal of Open Innovation, 5*(1), 18. https://doi.org/10.3390/joitmc5010018

Malle, Bertram F. (2016). Integrating robot ethics and machine morality: The study and design of moral competence in robots. *Ethics and Information Technology*, *18*(4), 243-256. DOI: https://doi.org/10.1007/s10676-015-9367-8

Marcus, Gary, & Davis, Ernest. (2020, August 22). GPT-3, Bloviator: OpenAI's language generator has no idea what it's talking about. *MIT Technology Review*. https://www.technologyreview.com/2020/08/22/1007539/gpt3-openai-language-generator-artificial-intelligence-ai-opinion/

McKee, Heidi A., & Porter, James E. (2017). AI agents as professional communicators. In *Professional Communication and Network Interaction: A Rhetorical and Ethical Approach*. New York: Routledge.

McKee, Heidi A., & Porter, James E. (2020). Ethics for AI writing: The importance of rhetorical context. *Proceedings of 2020 AAAI/ACM Conference on AI, Ethics, and Society* (AIES'20), February 7–8, 2020, New York, NY, USA. DOI: https://doi.org/10.1145/3375627.3375811

OpenAI. (2019). Better language models and their implications. https://openai.com/blog/better-language-models/

Radford, Alec, et al. (2019). Language models are unsupervised multitask learners. *OpenAI Blog*. https://openai.com/blog/better-language-models/

Russo, Federica. (2018). Digital technologies, ethical questions, and the need of an informational framework. *Philosophy & Technology, 31*(4), 655-667. DOI: https://doi.org/10.1007/s13347-018-0326-2