I Contain Multitudes: A Typology of Digital Doppelgängers

Forthcoming in the *American Journal of Bioethics*

William D'Alessandro
Department of Philosophy, William & Mary

Trenton W. Ford
Department of Data Science, William & Mary

Michael Yankoski
Department of Data Science, William & Mary

The need for a plurality of doppelgänger designs

Iglesias et al. argue that "some of the aims or ostensible goods of person-span expansion could plausibly be fulfilled in part by creating a digital doppelgänger"—that is, an AI system designed to reproduce someone's traits or advance their projects after their natural death (2024, 5). Here and elsewhere, Iglesias et al. write as though they expect the human-doppelgänger relationship to exhibit two noteworthy features. First, the relationship will be *one-one*, in the sense that a single doppelgänger system will be trained for each person who opts to receive one. Second, doppelgängers will be *general-purpose*, in the sense that a universal design template will suffice for realizing most person-span-extension aims.

This picture is likely too simple. We expect there will be compelling technological, social, intellectual and legal reasons to deploy a variety of doppelgänger systems. These systems will embody distinct sets of design desiderata, each matched to some subset of the core aims of person-span extension. And it will often be practical for several of a given person's doppelgängers to be in deployment simultaneously.

Let us briefly motivate these claims. Suppose the goal is to devise a doppelgänger system for a given human subject *H*. One possible approach is to train a single doppelgänger D_H on as large a sample as possible of *H*'s memories, beliefs, skills, plans, personality traits and so on. This system could then be deployed in any desired context, drawing on the full range of its imitative capabilities to provide maximally *H*-like performance. In particular, one might hope to rely on D_H's understanding of *H*'s values and preferences to minimize uncharacteristic or otherwise unwanted behavior. Iglesias et al. seem to have something like this in mind when they envision a "technologically and ethically sophisticated" system which "won't, for instance... share intimate text messages previously sent to your spouse with your students, boss, or children", but which "could still use these messages to generate similar intimate texts under appropriate conditions: say, if and only if it can verify it is your partner who is conversing with it" (9).

Appealing as this picture may be, such universal doppelgängers have important drawbacks. A first problem is *data security*. It will often be of utmost importance to ensure that sensitive information about a subject's life and work is strictly controlled. But a singular all-purpose doppelgänger would possess a wealth of such data, and its indiscriminate deployment in uncontrolled contexts presents a large attack surface to malicious actors. Like current large language models, doppelgänger systems may prove susceptible to jailbreaking attacks (Wei et al. 2023), or alternatively to deceptive or manipulative inputs designed to exploit a subject's psychological vulnerabilities. Indeed, a hostile party might learn much even by probing the boundaries of a knowledgeable doppelgänger's refusals.

A second issue is that it will often be desirable for a given person's doppelgängers to have *distinct capabilities*, *knowledge and dispositions* in different deployment contexts. For instance, family members may wish for their version of a deceased relative's doppelgänger to accept multimodal inputs (for sharing pictures of the grandchildren while chatting over breakfast, say), to display a homelier personality, and to recall intimate family moments, while a doppelgänger meant for controlled interaction with the public might call for quite different design parameters.

Finally, it will be useful for some person-span-extension purposes if one's doppelgänger can continue to *learn* and *develop* (and perhaps even change and forget) after one's death. Other such goals, meanwhile, are best served by doppelgängers that permanently retain one's original personality, plans and memories. (Fan et al.'s (2024) work on Comp-HuSim agents, to which two of the present authors contributed, is an example of the latter type of architecture.)

We therefore doubt the *one-one* and *general-purpose* assumptions. Doppelgängers are likely to take heterogeneous forms in response to the diverse goals of person-span extension (along with various extrinsic pressures). While the space for possible customizations is large, some packages of needs and constraints will be relatively common, and we expect standard design types to emerge in response. We sketch four such types below.

The family heirloom

In line with Iglesias et al.'s *Relational* and *Legacy* aims, a likely use case for a personal doppelgänger is that of the *family heirloom*. The purpose of this type of system is to carry forward the memories and persona of a departed family member for the benefit of their surviving relatives, descendants and other intimates. While a high-quality heirloom should represent a generous cross-section of the original person's mind, the faithful expression of personality and preservation of autobiographical memories are likely to be especially important.

Heirloom doppelgängers also call for distinctive constraints. Given the complexity of family dynamics, careful attention may need to be paid to the system's interactions with potential interlocutors: as Iglesias et al. note, a given manner, tone or conversation topic might be appropriate for a spouse but not for a child or cousin. Going further, one can imagine demand for adaptive measures that intelligently evolve as the family tree grows over time, among other capabilities for learning and change. (A spouse bereaved at a young age may wish for their partner's doppelgänger

to continue growing alongside them, for instance.) Designers will also have to consider how, if at all, to adjust, restrict or enhance an heirloom's behavior to protect or comfort its users.

The public legacy

High-profile individuals often maintain distinct public and private lives. A *public legacy* doppelgänger would serve as a noteworthy figure's enduring outward persona: maintained, perhaps, by the deceased's estate or a specialized foundation, and accessible via the internet or a museum installation.

A legacy doppelgänger would offer the experience of personal interaction with a historic leader, celebrity or intellectual to a wide audience. It might be called upon to act as the figurehead of an organization or as an emblematic representation of the ideals of the deceased. (Perhaps one lens through which to consider this type of doppelgänger can be found in Max Weber's "Charismatic Individual" (Weber 1978), upon whom the expectations and desires of the audience are often projected.) Given its limited role and broad security exposure, a legacy system will aim to convey the subject's public persona and thinking in their specialty area, but its access to autobiographical memories, trade or state secrets and other sensitive information should be strictly controlled.

The research archive

Primary-source accounts are invaluable to historians seeking to understand the past. The *research archive* doppelgänger would provide scholars and other inquirers with a rich source of firsthand knowledge about the subject's life and times. Such doppelgängers might be administered by a university or government library and accessible only to authorized parties.

An archive system armed with comprehensive autobiographical details would be useful in the case of a prominent subject likely to attract direct interest from historians. But there would presumably be much demand also for doppelgängers trained extensively on an ordinary person's memories of living in a particular era, working in a particular field, participating in a particular subculture or the like. Here the importance of capturing personal idiosyncrasies or conveying a preferred side of oneself would often be relatively small.

The project surrogate

The *project surrogate* doppelgänger aims at the continuation of one's personal or professional initiatives after death. These might be efforts one began in life but was unable to complete, or long-horizon projects demanding supervision beyond a natural lifespan. An adept surrogate would need robust access to one's project-relevant skills and knowledge, and presumably also to one's motivating beliefs, values and commitments, but likely not to intimate personal memories.

Many will wish to equip their surrogates with autonomous agentic capabilities, and such systems raise complex questions. Decisions will have to be made about whom to hold responsible for a surrogate's actions, whether such a system can own property or control resources in its own right, when and by whom a surrogate can be deactivated before achieving its goals, and so on. Employers may seek to protect and control key employees' surrogates, or even to develop such doppelgängers themselves to ensure the longevity of their workforce.

Conclusion

We contain multitudes. A single type of digital doppelgänger won't suffice for all we hope to be and do after death. Ethicists, technologists, policymakers and others must begin to consider how the various systems to come are best implemented and governed.

This panoply of doppelgängers has philosophical implications too. As Iglesias et al. suggest, we may have reason for egoistic concern about entities whose behavior and goals align with our own (that is, with "*D*-related beings" (7)). We've claimed that multiple doppelgängers exhibiting varying kinds and degrees of *D*-relatedness are more likely than a single, comprehensively *D*-related doppelgänger. If so, which of these personlike fragments is (most) us?

References

Fan, C., Tariq, Z., Bhuiyan, N., Yankoski, M., & Ford, T. (2024). Comp-HuSim: Persistent Digital Personality Simulation Platform. In *Proceedings of the ACM UMAP 2024*.

Iglesias, S., Earp, B. D., Voinea, C., Mann, S. P., Zahiu, A., Jecker, N. S., & Savulescu, J. (2024). Digital Doppelgängers and Lifespan Extension: What Matters? *The American Journal of Bioethics*, 1–16. https://doi.org/10.1080/15265161.2024.2416133

Weber, M. (1978). *Economy and Society: An Outline of Interpretive Sociology* (G. Roth & C. Wittich, Eds.). University of California Press.

Wei, A., Haghtalab, N., & Steinhardt, J. (2023). Jailbroken: How Does LLM Safety Training Fail?. In *NeurIPS '23: Proceedings of the 37th International Conference on Neural Information Processing Systems,* 80079–80110.