Digital Immunity: human consciousness as shield for A.I. harms

Keywords: integrated human-machine decision-making; integrative modeling; large-scale social experiments; reproducibility; downsides of social media.

Extended Abstract

Every technological revolution comes with inevitable downsides and side-effects. Three decades into the digital revolution, the dangers of the computational paradigm have manifested themselves in form of A.I.-based influences manipulating the free will of human minds. While our digital mind-extensions have led to a myriad of benefits, they have also led to a pandemic of mental health issues (especially among young people), misinformation affecting political processes, e-commerce manipulation exploiting class differences, and online judgementalism poisoning cultural processes, among many other threads to individual and social wellbeing (e.g. Allcott et al., 2020; Eyal, 2019; Lanier, 2018; Nahai, 2013; Nodder, 2013; Orlowski, 2020; Vogel et al., 2014; Ward, 2022). While policy regulations, technological innovations, and corporate responsibility are being adjusted to mitigate specific downsides, we ask if A.I. exerts an evolutionary pressure on the superiority of human minds. Does the human mind have to adapt by evolving a kind of 'digital immunity' to protect itself from tailor-made stimuli of our digital mind extensions?

We hypothesize that individuals with psychological traits typical for self-transcended individuals have already evolved a natural resistance to harms from A.I. driven persuasive technology, such as typical on online social media. We theorize that this 'digital immunity' is a manifestation of people's consciousness development when reaching the stage Maslow (1971) called 'self-transcendence'. Our general hypothesis is that psychological self-transcendence provides digital immunity because both originate from the de-conditioning of mental stimuli and response. This is in line with May's "Psychological Bases of Freedom", where he defined "mental health as the capacity to be aware of the gap between stimulus and response, together with the capacity to use this gap constructively" (May, 1962, p. 43).

Given the lack of theoretical guidance on the relationship of A.I. and human consciousness development, we pursue an integrative modeling approach that iterates between explanation and prediction (as proposed by the computational social science community, i.e. Hofman et al., 2017, 2021). See the Figure 1 for the general workflow of our integrative modeling framework. We execute three consecutive rounds of large-scale experimental surveys, with each consecutive round replicating the previous result through public pre-registration. Our first study (May 2021 – August 2022), casted a wide net in a 1.5h online experiment, collecting 29 measures on human consciousness and having participants going through 28 experiments in a mock-social media environment that simulated YouTube, Facebook, Instagram, Twitter, WhatsApp, Amazon, and eBay. We obtained 1,164 high quality responses from 2,144 participants.

Our integrative modeling framework that iteratively investigates exploratory (machine learning) and confirmatory (theoretical) approaches, produced a model that was able to explain over 80% of the variance of the negative correlation between four digital harms and our first 'digital immunity proxy', which mainly consisted of only three measures, namely mindfulness (FFMQ), self-compassion (SCS) and nonduality (NETI) (see Figure 2).

We preregistered a replication study (for more see authors…) in Sept. 2022 and successfully replicated the model from Figure 1 (RMSEA = 0.064, with <0.08 being the acceptable threshold in Structural Equation Models of such complexity), as well as three additional pre-registered hypotheses. The second study (Oct. 2022 – Mar. 2023) consisted of 1,277 high quality responses (from a total of 2,328 participants), who contributed to a more pin-pointed 35 min online experiment.

Informed by our previous findings, we introduced two new measures in Study 2 and explored their possible role (namely psychological flexibility/ Acceptance and Commitment Theory (ACT), and interoception (MAIA)). They turned out to be useful to inform digital immunity. We employed our integrative modeling workflow (see Figure 1) to explore more sophisticated hierarchical models (see Figure 3). In line with our initial hypothesis about the underlying generative mechanism of digital immunity, the shown hierarchical model of digital immunity suggests that '*consciousness states*', which include psychological skills, outlooks and habits, like mindfulness and self-compassion (which can be trained), are backed up by a more long-term '*consciousness stages*', which include psychological traits of human development, which often correlate with stable psychological traits, age and education (Weltanschauung), as they evolve over a lifetime. This two part structure suggests that some aspects of digital immunity might be trainable, such as through courses on positive psychology (Seligman & Csikszentmihalyi, 2000), but that the long-term solution for human minds to share mind-spaces with manipulative A.I. might consist in an evolution to higher stages of consciousness, as long postulated by the literature on self-transcendence (Sartre, 1957; Frankl, 1966; Maslow, 1971; O'Fallon, 2020).

We are currently executing a third study, again with public pre-registration, to dive deeper into the generative mechanism and the exploration of possible ways to foster much needed 'digital immunity' to the mental threats posed by quickly advancing artificial intelligence (A.I.) and its persuasive technology. This third round includes qualitative focus groups to complement our computational integrative modeling workflow.

References

Eyal, N. (2014). *Hooked: How to Build Habit-Forming Products*. Penguin.

Hofman, J. M., Sharma, A., & Watts, D. J. (2017). Prediction and explanation in social systems. *Science*, 355(6324), 486–488.

Hofman, J. M., Watts, … Vespignani, A., & Yarkoni, T. (2021). Integrating explanation and prediction in computational social science. *Nature*, 595(7866), 181–188.

Lanier, J. (2018). *Ten Arguments for Deleting Your Social Media Accounts Right Now*. Henry Holt and Company.

Maslow, A. H. (1971). *The farther reaches of human nature*. Arkana/Penguin Books.

O'Fallon, T. (2020). States and STAGES: Waking up Developmentally. *Integral Review: A Transdisciplinary & Transcultural Journal for New Thought, Research, & Praxis*, 16(1).

Sartre, J.-P. (1957). *The Transcendence of the Ego: An Existentialist Theory of Consciousnes*s. Macmillan.

Seligman, M. E. P., & Csikszentmihalyi, M. (2000). Positive psychology: An introduction. *American Psychologist*, 55(1), 5–14.

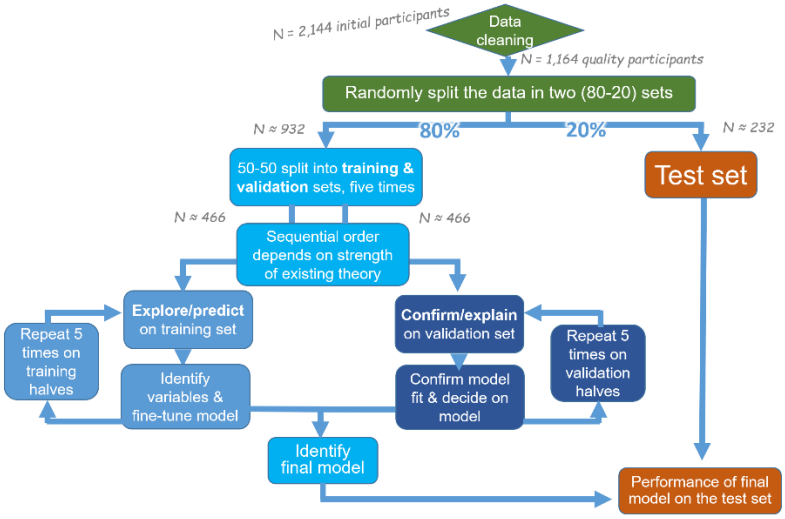

Figure 1. Integrative Modeling Framework developed for our study, with examples from Study 1

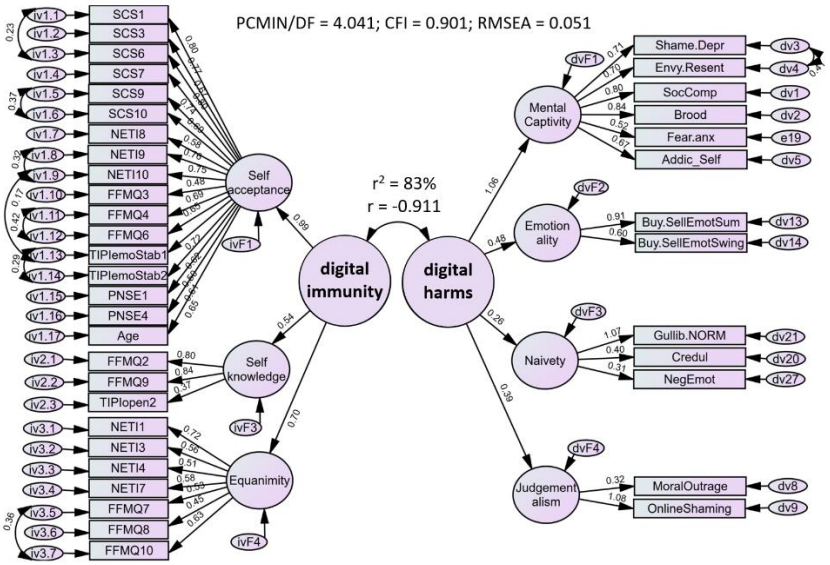

Figure 2. Final superior model identified by Study 1; successfully replicated in preregistered replication Study 2.

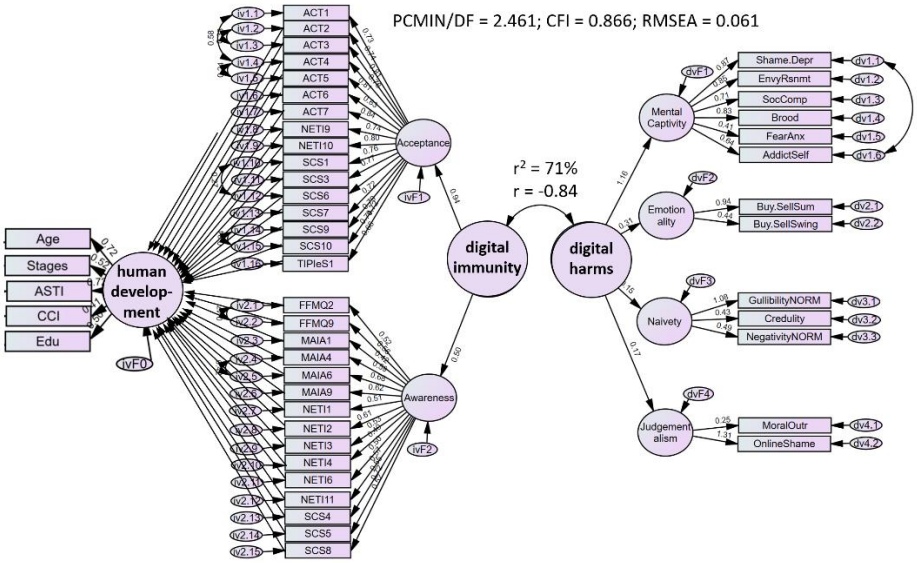

Figure 3. Hierarchical model resulting from integrative modeling of Replication Study 2.