Kelly Quinn

BIOETHIC 505

March 2021­­­­­­­­

**Ethical Engineering: Computational Persuasion for Positive Behavior Change**

In the past few years both private and public sector companies and agencies have entered into discussion about or issued guidelines for ethics in artificial intelligence. While the global consensus agrees that ethics in artificial intelligence is of paramount importance, the notion of what exactly ethical artificial intelligence would look like remains unsatisfactorily defined. International organizations and corporations such as Google, SAP, Amnesty International, and the High-Level Expert Group on Artificial Intelligence as appointed by the European Union, to name a few, have been tasked or taken it upon themselves to draft such guidelines (Jobin et al., 2019). In many instances, these bodies have turned to classic theories of ethics and morality from Aristotle’s teleological arguments to Kant’s deontological (Chatila & Havens, 2019), but just as these historical figures could not have predicted this great dilemma of the 21st century, modern society is struggling to cope with the speed of innovation and technological advancement. One area in particular that is ripe for ethical dissertation is the development of persuasive technology and its uses on the influence the behavior and attitudes of the 21st century human being. Even when it may be for the greater good.

*Computational Persuasion*

The dialogue on artificial intelligence (AI) oscillates between enthusiasm and fear. Fear remains strong that AI may render human capital and labor obsolete, become inadvertently programmed with human biases, or be used globally by bad actors to do harm. While there is general consensus on particular ethical domains in which AI technology falls (Jobin et al., 2019) this paper will explore transparency, non-maleficence, and beneficence. Transparency remains the most discussed ethical topic in a literature search undertaken by Chatila and Havens (2019) but it is consistently discussed as a way to combat harm potential (Jobin et al., 2019). The concept of transparency in AI typically refers to the ‘discoverability’ of the AI’s decision-making strategies, processes, or intention. Furthermore, this should be available and accessible to a wide range of stakeholders, from investors to the average user. Key concepts to consider for transparency are: the ability to trace the process to the source, the ability to explain the process or decision-making rationale, and how easy it is to interpret (Chatila & Havens, 2019).

Non-maleficence, or the concept that AI should cause neither intention nor unintentional harm, is more commonly referenced than the positive valence of beneficence. To clarify, not only should the AI do no harm, but it should go a step further and promote good. Chatila and Havens (2019) list notable examples to include the promotion of human well-being, peace, happiness, elevation of socio-economic improvement or the creation of opportunity. To state it broadly, well-being encompasses the entirety of the human experience and the internal and external forces upon which life depends. The internal and external factors in social sciences are often depicted as the socio-ecological model and range from individual through macro-societal impacts. The concept of well-being is an oft mentioned but seldomly unpacked “catch all” in social sciences. There are a variety of measurable markers that have been developed to determine level of well-being across the ecological model. For an individual stakeholder in AI interaction, these effects may be observed in their psychological or physical well-being markers (Chatila & Havens, 2019).

The reality of AI development however is dependent on human engineering and may therefore never truly be neutral (Oinas-Kukkonen, 2013). These technologies were developed with a purpose or a goal and are thus, arguably, never ‘switched off.’ As app developers aim to remain relevant and public sector agencies leverage AI to solve ‘wicked problems’ (Rittel & Webber, 1973), they have turned to psychology to understand how to use influence and persuasion to nudge (Thaler & Sunstein, 2009) their users into making the desired decision. While persuasion is not in itself negative and influence is a natural course of social interaction, there is ethical consideration on the nature of the persuasion, the intention of the persuader, the susceptibility of the target, and the potential that any such influence was actually an accidental side-effect (Oinas-Kukkonen, 2013).

This dive into the engineering of influence has opened the door for a new type of persuasion, computational persuasion (Hunter, 2018). Computational persuasion is ‘the study of formal models of dialogues involving arguments and counterarguments, of persuasion models, and strategies for APSs [automated persuasion system]” (Hunter, 2018). An important distinction made by Hunter (2018) is that these models are developed on the expansion of computational models of argument. They are not meant to replicate or replace models for human persuasion, but rather for the production of models that are usable by artificial intelligent systems to successfully persuade a human subject.

Recent studies have shown that persuasive models that are comparatively successful are more effective in influencing behavior when they are personalized to the target (Matz et al., 2017). Typical social science behavior measures are taken through large scale surveys and questionnaires, which open a host of bias issues (Kreitchmann et al., 2019). However, recent field research in computation science is finding that elements of psychological profiles may actually be predictable based on a user’s ‘digital footprint.’ This digital footprint is the algorithmic collective of measurable online actions such as Tweets, Facebook Likes, Instagram posts, blog posts, and advertisement engagements (Matz et al., 2017). What then, might this mean for the realm of behavior change?

*Persuasion and Behavior Change*

Persuasion is a common activity that occurs in social interaction, defined as one party attempting to induce another party into a desired behavior or belief. As AI interaction becomes more commonplace in daily life, so too are computational solutions being leveraged to solve problems. A verdant research ground in the intersection of computational persuasion and social sciences is its potential to intervene in behavior change (Hunter, 2018). The six fundamental domains of persuasion as spearheaded by Cialdini of reciprocation, consistency, social proof, liking, authority, and scarcity (Cialdini, 1984) are finding roots in the new age of computational persuasion. However, because interaction with technology is not naturally occurring, there have been challenges that AI has had to cope with, which have been guided by psychological principles. Particularly, that persuasive technologies for behavior change should be both useful and easy to use, that they should be unobtrusive to the user’s primary objective, and as discussed previously, the persuasion technology should be transparent (Bezuidenhout & Ratti, 2020). It remains a challenge for persuasive technology in the domains of behavior change to communicate messages that are clear and comprehensible and also effectively persuade the target (Dragoni et al., 2020) without dipping into questionable ethics.

Domains within behavior change for healthy outcomes that are well suited for persuasive technology include healthy lifestyle adoptions, addiction management, dietary and obesity management, chronic disease management or treatment compliance (Hunter, 2018), risky sexual behavior and unplanned pregnancy (Orji & Moffatt, 2018). There are vast theoretical underpinnings to justify computational persuasion as effective for behavior change (Oinas-Kukkonen, 2013). The Theory of Reasoned Action (Fishbein & Ajzen, 1975), an individual’s attitude toward a behavior, and Theory of Planned Behavior (Ajzen, 1991), an individual’s perception of ease regarding performing the behavior, fit quite nicely with the Technology Acceptance Model (Davis, 1989), an individuals perceived ease of use of a system and the system’s usefulness. Interactive models such as Fishbein’s (Fishbein, 2000) take composites of multiple variables from diverse behavioral science theoretical models to determine the likelihood of an individual’s intention to use a system based on perceived ease of use, strength of performance intention, and belief in one’s own skills or efficacy (Davis, 1989). There is further evidence that persuasive technology dovetails with behavior change models by looking at the complementary nature of APSs and The States of Change Model (Prochaska & Velicer, 1997). The States of Change Model is comprised of five ‘states’ an individual may experience regarding a behavior or attitude change: pre-contemplation, contemplation, preparation, action, and maintenance (Hunter, 2018). Hunter (2018) argues that if an APS could begin a persuasive intervention at the contemplation state, the user may be amenable to interacting with an AI regarding the behavior change.

*An Ethical Dilemma*

Persuasion, rather than coercion or negative reinforcement, is considered as a more effective and ethical approach to achieving a desired and sustained behavior change (Kight & Gram-Hansen, 2019). In the field of psychology, social science, and computation the topic of persuasion remains coupled with the concept of ethics. Fogg (Fogg, 2003) in his seminal research emphasized ethics as a defining feature of persuasion. Persuasion is not normative according to Hunter (2018), meaning that there are no steadfast principles to persuasion. Persuasive arguments can be inconsistent, untrue, biased, irrational, and still manage to persuade a target. However, knowing this to be true raises a host of ethical questions. This is where regulatory agencies may find their role in preventing bad actors from exploiting the masses to the manipulative side of persuasion and ‘surveillance capitalism’ (Zuboff, 2019). It is universally agreed that coercion and deception are unethical. In fact, utilitarian approaches to behavior change support systems are designed to form and reinforce behavior change without using coercion or deception (Oinas-Kukkonen, 2013), yet, surveillance and manipulation may fall into a grey area in which their ethical merit is contextual. Atkinson (2006) argues that persuasive technology may be ethical in situations in which the user is unaware of the programmers’ intent (lack of transparency) if the behavioral change was one of beneficence. She goes on, however, to pose the question of whether or not it may be more ethically sound if such benevolence was imparted through education and advocacy (Kight & Gram-Hansen, 2019).

*Conclusion*

As the inevitable sea of digitalization advances, leaving little option to disengage and none to opt out, the ethical issues surrounding AI interaction are commonplace. Humankind is forced to grapple with the construction, utilization, and new realities of a digital environment (Bezuidenhout & Ratti, 2020). As computational persuasion is interwoven into every digital interaction and large data companies can sell your digital footprint to the highest bidder, remaining untouched by artificial intelligence systems and their attendant influence seems increasingly impossible. Likewise, expecting the average user to expect, demand, and comprehend full transparency may also seem like a pipedream in a society that touts a motto of ignorance as bliss. However, the stakes are impossibly high. These past few years, the world sat front row to one of the greatest examples of digital influence by way of social media in the rise of Donald Trump’s presidency. As the public sector lags behind in AI integration, adoption, and regulation an ethical question for a parting thought remains, whose role is it to protect the masses from themselves?

**Resources**

Atkinson, B. M. C. (2006). Captology: A Critical Review. In W. A. IJsselsteijn, Y. A. W. de Kort, C. Midden, B. Eggen, & E. van den Hoven (Eds.), *Persuasive Technology* (pp. 171–182). Springer. https://doi.org/10.1007/11755494\_25

Bezuidenhout, L., & Ratti, E. (2020). What does it mean to embed ethics in data science? An integrative approach based on microethics and virtues. *AI & SOCIETY*. https://doi.org/10.1007/s00146-020-01112-w

Chatila, R., & Havens, J. C. (2019). The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. In M. I. Aldinhas Ferreira, J. Silva Sequeira, G. Singh Virk, M. O. Tokhi, & E. E. Kadar (Eds.), *Robotics and Well-Being* (Vol. 95, pp. 11–16). Springer International Publishing. https://doi.org/10.1007/978-3-030-12524-0\_2

Cialdini, R. (1984). *Influence: The Psychology of Persuasion*. HarperCollins.

Davis, F. D. (1989). Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology. *MIS Quarterly*, *13*(3), 319–340. https://doi.org/10.2307/249008

Dragoni, M., Donadello, I., & Eccher, C. (2020). Explainable AI meets persuasiveness: Translating reasoning results into behavioral change advice. *Artificial Intelligence in Medicine*, *105*, 101840. https://doi.org/10.1016/j.artmed.2020.101840

Fishbein, M. (2000). The role of theory in HIV prevention. *AIDS Care*, *12*(3), 273–278. https://doi.org/10.1080/09540120050042918

Fogg, B. J. (2002). *Persuasive Technology: Using Computers to Change What We Think and Do.* Elsevier. http://www.123library.org/book\_details/?id=34414

Hunter, A. (2018). Towards a framework for computational persuasion with applications in behaviour change. *Argument & Computation*, *9*(1), 15–40. https://doi.org/10.3233/AAC-170032

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, *1*(9), 389–399. https://doi.org/10.1038/s42256-019-0088-2

Kight, R., & Gram-Hansen, S. B. (2019). Do Ethics Matter in Persuasive Technology? In H. Oinas-Kukkonen, K. T. Win, E. Karapanos, P. Karppinen, & E. Kyza (Eds.), *Persuasive Technology: Development of Persuasive and Behavior Change Support Systems* (pp. 143–155). Springer International Publishing. https://doi.org/10.1007/978-3-030-17287-9\_12

Kreitchmann, R. S., Abad, F. J., Ponsoda, V., Nieto, M. D., & Morillo, D. (2019). Controlling for Response Biases in Self-Report Scales: Forced-Choice vs. Psychometric Modeling of Likert Items. *Frontiers in Psychology*, *10*. https://doi.org/10.3389/fpsyg.2019.02309

Matz, S. C., Kosinski, M., Nave, G., & Stillwell, D. J. (2017). Psychological targeting as an effective approach to digital mass persuasion. *Proceedings of the National Academy of Sciences*, *114*(48), 12714–12719. https://doi.org/10.1073/pnas.1710966114

Oinas-Kukkonen, H. (2013). A foundation for the study of behavior change support systems. *Personal and Ubiquitous Computing*, *17*(6), 1223–1235. https://doi.org/10.1007/s00779-012-0591-5

Orji, R., & Moffatt, K. (2018). Persuasive technology for health and wellness: State-of-the-art and emerging trends. *Health Informatics Journal*, *24*(1), 66–91. https://doi.org/10.1177/1460458216650979

Prochaska, J. O., & Velicer, W. F. (1997). The transtheoretical model of health behavior change. *American Journal of Health Promotion: AJHP*, *12*(1), 38–48. https://doi.org/10.4278/0890-1171-12.1.38

Rittel, H. W. J., & Webber, M. M. (1973). Dilemmas in a general theory of planning. *Policy Sciences*, *4*(2), 155–169. https://doi.org/10.1007/BF01405730

Thaler, R. H., & Sunstein, C. R. (2009). *Nudge: Improving decisions about health, wealth, and happiness* (Rev. and expanded ed). Penguin Books.

Zuboff, S. (2019). *The age of surveillance capitalism: The fight for a human future at the new frontier of power* (First edition). PublicAffairs.