Safety From Ethical Hazards

Prospects for a contribution from software engineering

Kurt C. Wallnau
Software Engineering Institute, Carnegie Mellon University
Pittsburgh, PA, USA
kcw@sei.cmu.edu

ABSTRACT

In this paper, I argue that while normative ethical concerns such as fairness and accountability must be addressed in the design of intelligent software, these concerns are far removed from traditional software engineering practice. After reviewing representative illustrations of such ethical hazards arising from the use of intelligent applications, I discuss current practices that might contribute to software engineering practices to assure that intelligent software is safe to use, i.e., free of ethical hazard.

CCS CONCEPTS

• **Human-centered computing → Human computer interaction (HCI)**; • **Social and professional topics**; • **Software and its engineering**;

KEYWORDS

Algorithmic intelligence, smart applications, cognitive services, psychosocial risk

ACM Reference Format:
Kurt C. Wallnau. 2018. Safety From Ethical Hazards: Prospects for a contribution from software engineering. In *SE4COG'18: SE4COG'18:IEEE/ACM 1st International Workshop on Software Engineering for Cognitive Services , May 28–29, 2018, Gothenburg, Sweden.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3195555.3195569

1 INTRODUCTION

In the past few years, we have witnessed a small number of breathtaking showcases of ML technology. More significant still are the myriad applications of ML and big data analytics. These have become a pervasive, if concealed, determinant of online *and* offline social reality—in mediating how we interact with one another, with businesses, and with government. And this has happened over a remarkably short period of time.

In this paper I refer to the various forms of ML and data analysis technology as *algorithmic intelligence*, and applications/services that make use of these enablers as intelligent applications/services, or sometimes simply intelligent software. There is now a substantial body of knowledge and practice available to software developers about various forms of algorithmic intelligence, and how they are used. Open source software libraries for building intelligent applications are widely available, and commercial platforms are evolving quickly in both the capabilities offered and the ease with which these capabilities can be incorporated into new applications. Commercial competition for expertise and mindshare is fierce. Colleges and universities are racing to meet industry demand by incorporating elements of machine learning, data mining and big data into their undergraduate curricula.

As the software industry has learned about the many ways that intelligent software can be used, it has also become increasingly aware of its ethical hazards. Using the framework proposed by Mittlestadt, et al [18], I distinguish epistemic concerns from normative concerns. I argue that epistemic concerns, however challenging they may be, are comprehended by the technical languages of software engineering discourse, namely mathematics and logic. However, normative concerns are value-laded, and arise from the social-embeddedness of decision outcomes. The language of discourse of normative ethics includes moral concepts such as fairness and agency that are not within the traditional ken of software developers. As such, normative concerns are not likely to be adequately addressed by software developers, and therefore will, without appreciable attention, continue to be a source of risk.

2 ETHICAL HAZARDS OF ALGORITHMIC INTELLIGENCE

As intelligent software has become increasingly present in our lives, we have also become acquainted with unique and unexpected hazards it poses to society. Algorithmic bias is a frequently-cited normative ethical concern. It takes many forms, for example bias against gender [4, 5, 8, 15, 17], sexual orientation [14], and race [1, 2, 28].

For one concrete example, gender was shown to affect job-related advertisements appearing in browsers[7], with male users receiving more advertisements than females for career coaching for high paying jobs. Face recognition applications have encountered difficulties when applied to images of people with dark skin [6]. Two widely-used tagging systems have been shown to tag black people's faces as apes and gorillas [13, 20]. COMPAS is a risk assessment software that has been widely used in criminal justice systems in the United States to predict recidivism. When assessing the risk, the software correctly predicted recidivism for black and white defendants, at comparable rates. However, racial bias was revealed where the predictions were erroneous, with black defendants twice as likely as whites to be labeled as high-risk but later found to not re-offend; conversely, white defendants labeled as low-risk were twice as likely to re-offend [26].

SE4COG'18, May 28–29, 2018, Gothenburg, Sweden

ACM ISBN 978-1-4503-5740-1/18/05…$15.00
https://doi.org/10.1145/3195555.3195569

Intelligent applications have also been shown to produce unexpected hazards to individual mental health [16, 23, 30], and to sense of autonomy [18]. Personalization algorithms risk imposing outcomes that give preference to third-party interests rather than those of the individual [3, 27]. Systems intended to reduce a user's experience of "information overload" have been shown to unintentionally manipulate the user's emotional state [16]. Emotional coercion is an intentional feature of free-to-play (F2P) online games, with games adaptively responding to user actions for the purpose of maximizing the likelihood of in-game purchases [12]. These F2P monetization strategies have been shown to be highly correlated with internet gaming disorder, a type of mental disorder manifested in loss of control, withdrawal, and clinically significant impairment and distress [10, 11].

3 SOFTWARE ENGINEERING RESPONSES TO NORMATIVE ETHICAL CONCERNS?

We use our laptop computers and cell phones with tacit and (I claim) reflexive expectation that they will not, for example: explode, burst into flames, shatter into thousands of razor-sharp shards, or emit toxic fumes. Violations do occur, but rarely; and when they do, they are newsworthy. The same tacit expectation of safety attends virtually all of the manufactured, functional things we regularly use. This is in its own way as remarkable an outcome as any showcase demonstration of machine learning; and this outcome can be attributed to *routine engineering practice*. All engineering professions codify ethical concerns that refer, in some way, to public safety and public welfare.

As we have seen, intelligent software poses many normative ethical hazards. An urgent question is whether, and how, software engineering practice can provide assurance that intelligent software is free from ethical hazard, and in particular free from normative hazard.

Algorithmic transparency has been suggested as a way to address normative ethical concerns. For example, there are various approaches to providing transparency and feedback about algorithmic state to enable inspection and steering of outcomes [9, 19]. Some have suggested that auditing will be a necessary step to verify correct function for all types of algorithmic intelligence [18]. Auditing is to create *ex post procedural* records of complex ML algorithms to deconstruct problematic, inaccurate decisions, or to detect discrimination or similar harms. A different notion of audit was proposed by Varshney, who suggests that procedural safeguards such as use of auditable open source software are needed to increase the safety of ML [29]. These solutions are all difficult to achieve with the currently generation of machine learning technology.

It is also possible to avoid hazards *a priori* by intentionally debiasing input data or introducing discrimination-defeating features into the ML classifiers [21]. These, in effect, prescribe judgments of fairness or desired social outcomes. This begs the question of who it is that prescribes these desired outcomes? One might be the established framework of discrimination law; although unrelated to discrimination law, legal doctrine for algorithmic contract law is already emerging [22]. Another, less formal approach is to invite vulnerable communities and independent advocacy groups to engage in discussion of fairness or other normative concerns [25].

4 CONCLUSION

I have argued that significant attention is needed to ensure that intelligent applications are free of normative ethical hazards. The extent to which this response resides within software engineering practice remains an open question. On the one hand, enhancements of machine learning architecture to improve transparency and traceability would seem to fall within the provenance of software engineering. On the other hand, a purely technical focus on transparency does not address safety, at least in the sense that safety is concerned with preventing hazards.

The ML genie is out of the bottle, and the consequences of this are still unfolding. An encouraging recent development is taking place in classrooms where the next generation of software engineers are being trained. Harvard University and Massachusetts Institute of Technology have introduced a course on the ethics and regulation of artificial intelligence, and the University of Texas at Austin a course on ethical foundations of computer science [24]. The true potential for smart applications will be obtained only when the technology is responsive to the ethical demands of society.

ACKNOWLEDGMENTS

The author thanks Kai Wei for her valuable insights. This material is based upon work funded and supported by the Department of Defense under Contract No. FA8702-15-D-0002 with Carnegie Mellon University for the operation of the Software Engineering Institute, a federally funded research and development center. DM18-0386

REFERENCES

[1] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. *ProPublica, May* 23 (2016).

[2] Julia Angwin and Terry Parris Jr. 2016. Facebook lets advertisers exclude users by race. *ProPublica blog* 28 (2016).

[3] Sally A Applin and Michael D Fischer. 2015. New technologies and mixed-use convergence: How humans and algorithms are adapting to each other. In *Technology and Society (ISTAS), 2015 IEEE International Symposium on*. IEEE, 1–6.

[4] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems*. 4349–4357.

[5] Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science* 356, 6334 (2017), 183–186.

[6] Christina Couch. 2017. Ghosts in the Machine. *PBS* (2017).

[7] Amit Datta, Anupam Datta, Jael Makagon, Deirdre K Mulligan, and Michael Carl Tschantz. 2018. Discrimination in Online Advertising: A Multidisciplinary Inquiry. In *Conference on Fairness, Accountability and Transparency*. 20–34.

[8] Amit Datta, Michael Carl Tschantz, and Anupam Datta. 2015. Automated experiments on ad privacy settings. *Proceedings on Privacy Enhancing Technologies* 2015, 1 (2015), 92–112.

[9] Nicholas Diakopoulos. 2016. Accountability in algorithmic decision making. *Commun. ACM* 59, 2 (2016), 56–62.

[10] M Dreier, K Wölfling, E Duven, S Giralt, ME Beutel, and KW Müller. 2017. Free-to-play: about addicted Whales, at risk Dolphins and healthy Minnows. Monetarization design and internet gaming disorder. *Addictive behaviors* 64 (2017), 328–333.

[11] Wendy Feng, Danielle E Ramo, Steven R Chan, and James A Bourgeois. 2017. Internet gaming disorder: Trends in prevalence 1998-2016. *Addictive behaviors* 75 (2017), 17–24.

[12] Robert Flunger, Andreas Mladenow, and Christine Strauss. 2017. The Free-to-play Business Model. In *The 19th International Conference on Information Integration and Web-based Applications and Services.*

[13] The Guardian. 2015. Flickr faces complaints over 'offensive' auto-tagging for photos. *Available: https://www.theguardian.com/technology/2015/may/20/flickr-complaints-offensive-auto-tagging-photos* (2015).

[14] Saikat Guha, Bin Cheng, and Paul Francis. 2010. Challenges in measuring online advertising systems. In *Proceedings of the 10th ACM SIGCOMM conference on*

Internet measurement. ACM, 81–87.

[15] Matthew Kay, Cynthia Matuszek, and Sean A Munson. 2015. Unequal representation and gender stereotypes in image search results for occupations. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 3819–3828.

[16] Adam DI Kramer, Jamie E Guillory, and Jeffrey T Hancock. 2014. Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences* 111, 24 (2014), 8788–8790.

[17] Anja Lambrecht and Catherine E Tucker. 2016. Algorithmic bias? An empirical study into apparent gender-based discrimination in the display of STEM career ads. (2016).

[18] Brent Daniel Mittelstadt, Patrick Allo, Mariarosaria Taddeo, Sandra Wachter, and Luciano Floridi. 2016. The ethics of algorithms: Mapping the debate. *Big Data & Society* 3, 2 (2016), 2053951716679679.

[19] Thomas Mühlbacher, Harald Piringer, Samuel Gratzl, Michael Sedlmair, and Marc Streit. 2014. Opening the black box: Strategies for increased user involvement in existing algorithm implementations. *IEEE transactions on visualization and computer graphics* 20, 12 (2014), 1643–1652.

[20] Molly Mulshine. 2015. A major flaw in Google's algorithm allegedly tagged two black people's faces with the word 'gorillas'. *Available: http://www.businessinsider.com/google-tags-black-people-as-gorillas-2015-7* (2015).

[21] Andrea Romei and Salvatore Ruggieri. 2014. A multidisciplinary survey on discrimination analysis. *The Knowledge Engineering Review* 29, 5 (2014), 582–638.

[22] Lauren Henry Scholz. 2016. Algorithmic contracts. (2016).

[23] Holly B Shakya and Nicholas A Christakis. 2017. Association of Facebook use with compromised well-being: a longitudinal study. *American journal of epidemiology* 185, 3 (2017), 203–211.

[24] Nastasha Singer. 2018. Universities Rush to Roll Out Computer Science Ethics Courses. *New York Times* (2018).

[25] Michael Skirpan and Micha Gorelick. 2017. The Authority of" Fair" in Machine Learning. *arXiv preprint arXiv:1706.09976* (2017).

[26] Matthias Spielkamp. 2017. Inspecting Algorithms for Bias. *MIT Technology Review* (2017).

[27] Meredith Stark and Joseph J Fins. 2013. Engineering medical decisions: computer algorithms and the manipulation of choice. *Cambridge Quarterly of Healthcare Ethics* 22, 4 (2013), 373–381.

[28] Latanya Sweeney. 2013. Discrimination in online ad delivery. *Queue* 11, 3 (2013), 10.

[29] Kush R Varshney. 2016. Engineering safety in machine learning. In *Information Theory and Applications Workshop (ITA), 2016*. IEEE, 1–5.

[30] Heather Cleland Woods and Holly Scott. 2016. # Sleepyteens: social media use in adolescence is associated with poor sleep quality, anxiety, depression and low self-esteem. *Journal of adolescence* 51 (2016), 41–49.