Not in my AI: Moral engagement and disengagement in health care AI development*

Ariadne A. Nichol, Meghan C. Halley, Carole A. Federico*, and Mildred K. Cho†
Stanford Center for Biomedical Ethics, Stanford University
Stanford, CA 94305, USA
Email: micho@stanford.edu

Pamela L. Sankar†
Department of Medical Ethics & Health Policy, University of Pennsylvania
Philadelphia, PA 19104, USA

Machine learning predictive analytics (MLPA) are utilized increasingly in health care, but can pose harms to patients, clinicians, health systems, and the public. The dynamic nature of this technology creates unique challenges to evaluating safety and efficacy and minimizing harms. In response, regulators have proposed an approach that would shift more responsibility to MLPA developers for mitigating potential harms. To be effective, this approach requires MLPA developers to recognize, accept, and act on responsibility for mitigating harms. In interviews of 40 MLPA developers of health care applications in the United States, we found that a subset of ML developers made statements reflecting moral disengagement, representing several different potential rationales that could create distance between personal accountability and harms. However, we also found a different subset of ML developers who expressed recognition of their role in creating potential hazards, the moral weight of their design decisions, and a sense of responsibility for mitigating harms. We also found evidence of moral conflict and uncertainty about responsibility for averting harms as an individual developer working in a company. These findings suggest possible facilitators and barriers to the development of ethical ML that could act through encouragement of moral engagement or discouragement of moral disengagement. Regulatory approaches that depend on the ability of ML developers to recognize, accept, and act on responsibility for mitigating harms might have limited success without education and guidance for ML developers about the extent of their responsibilities and how to implement them.

Keywords: Machine learning, Moral disengagement, Moral awareness, Regulation, Safety, Ethics

* C.A.F. was supported on a training grant from the National Institutes of Health (T32 HG008953).

† This work was supported by grants from The Greenwall Foundation and the National Institutes of Health (R01HG010476)

1. Introduction

Machine learning (ML) is increasingly utilized in health care, but can pose a variety of harms and raise ethical concerns. (Chen et al., 2021) Yet, unique features of ML create challenges to evaluating its safety and efficacy and minimizing harms. (Char, Shah, & Magnus, 2018 ; London, 2022) Proposed regulatory approaches designed to meet these challenges would shift the locus of responsibility for assessing and mitigating potential harms to ML developers. (US Food & Drug Administration, 2018, 2019, 2021) The success of such an approach would depend on the ability of ML developers to recognize, accept, and act on responsibility for mitigating harms. Other research suggests that the environment of computer science and software development could contribute to deflection of responsibility for harms (Gotterbarn, 2001; Vakkuri, Kemell, Jantunen, & Abrahamsson, 2020) in ways that are at odds with the culture of health care. We previously found that developers or machine learning-based predictive analytics for health care (MLPA) recognized a wide range of potential harms to individuals, social groups, and to the health care system. (Nichol, 2022) In addition, some developers were able to identify drivers of these harms and strategies to respond to these drivers through the development process. Those findings suggested that some MLPA developers acknowledge harms of their products and recognize strategies to mitigate those harms. However, recognition of the potential for harms and their mitigators is insufficient to prevent manifestation of harms if developers do not have moral awareness – the appreciation that there is an ethical aspect to the decisions that *they* make. According to the Rest Model, there are four components of ethical decision-making: (1) moral awareness, (2) moral judgment, (3) moral intention, and (4) moral action. (Narvaez & Rest, 1995) That is, developers would, at the very least, have to accept responsibility for identifying and minimizing harms as a prerequisite for taking appropriate action. We present a new analysis of previously-collected data from interviews of health care MLPA developers in the US (Nichol, 2022) which examines developers' perceptions of moral awareness and responsibility.

2. Methods

2.1. *Recruitment*

We recruited individuals from July 2019 to July 2020 who were working for U.S.-based organizations involved in developing MLPA tools for use in health care settings. We selected individual organizations based on our previously published analysis of the landscape of predictive analytics in health care (Nichol et al., 2021) which included a range of organizational types and sizes. The sample consisted of computer software and information technology companies, including those specifically focused on health care, as well as health insurers and hospital systems. In addition, we classified organizations by size based on number of employees (1-50, 51-1,000, over 1,000), as specified in the LinkedIn page for each organization. We identified 96

organizations, of which we selected 15 that were representative of the range of organizations, both in terms of type and size. (Table 1)

From these organizations, we identified potential participants through LinkedIn, reviewing search results by organization for key words such as data scientist, software engineer, or manager. We contacted individuals to participate through LinkedIn's direct messaging feature. To identify additional participants, we also used a snowball sampling approach. (Bernard, 2006) To examine the MLPA development process from different perspectives, we intentionally included participants representing a variety of roles, including data scientists, software engineers, project managers and executive leaders, among others. Individuals were offered a $100 electronic gift card for participation. Our study was approved by the Institutional Review Boards of Stanford University and the University of Pennsylvania.

2.2. *Data collection*

Each participant completed a one-hour semi-structured interview through video conference. Interviews were conducted by one of two members of the research team (AAN or MCH). We iteratively developed the interview guide through pilot interviews with MLPA developers with familiarity with health care ML, and who were not included in the final sample. The interview guide included questions on the participants' background and training, company and MLPA product goals in health care, facilitators and barriers to product development, potential benefits and harms of these products, and views on their regulation and oversight.

2.3. *Data analysis*

Interviews were audio-recorded, transcribed verbatim and de-identified. We analyzed the data using the mixed-method analytic software Dedoose™ 8.3, using standard qualitative data analysis methods (Miles, Huberman, & Saldana, 2019) based on grounded theory. (Strauss & Corbin, 1997) To generate the initial codebook, all team members reviewed a subset of interview transcripts and generated a list of key concepts identified in the data. The team then iteratively refined the codebook through multiple rounds of provisional coding. Once the codebook was finalized, at least two team members independently coded each interview to enhance rigor and reliability, resolving any coding differences through team consensus. To further examine participant perceptions of the potential harms of MLPA in health care, and their attitudes toward regulation and oversight, we then reviewed all data coded to these topics across all participants to identify consistency and variability in narratives both within and across participants.

3. Results

3.1. *Participant characteristics*

40 of 76 MLPA developers contacted responded (52.6%). The majority (n=29, 72.5%) of participants worked at health care-oriented computer software and information technology companies. Almost two thirds (n=25) of participants held roles that involved both working directly

with data in MLPA development and other functions, such as leadership. Sixteen participants occupied high-level management roles. Thirty-five percent held health-related advanced degrees.

3.2. *Developer perspectives on responsibility*

In analyzing our interview data on developer perspectives on potential harms and benefits of their products, we found statements revealing their perspectives on roles and responsibilities to mitigate harms even though we did not ask about them directly. Some respondents indicated a sense of moral sensitivity or awareness that included recognition of moral issues and empathy with others' points of view, (Narvaez & Rest, 1995) and some of those reflected recognition of the developer's role in addressing these issues. Others made statements that minimized harmful impacts of their products or their responsibilities to mitigate them. Examples of these statements are described below.

3.3. *Moral disengagement*

Many developers made statements recognizing the potential harms from use of ML in health care, especially to patients, such as bias, loss of privacy, or inaccurate output of models. However, a subset of these statements also indicated minimization of harms or deflection of responsibility for preventing or mitigating them. We identified eight different subtypes of "moral disengagement" statements that created moral distance between their actions and harms or responsibility. (Table 2) These eight types of moral distancing or disengagement could be grouped into two categories: (1) rationalizations for, or minimization of harms of AI in health care applications (minimizing risk), or (2) minimization of the developer's role in addressing or mitigating harms (minimizing responsibility).

Examples of each of the eight subtypes of moral disengagement statements are shown in Table 2, and the label we gave to each subtype. Some of these statements compared the harms of ML to those in other contexts such as social media, or financial data and asserted that there was no difference between those contexts and health care (Subtype: ***No difference***). Others favorably compared the harms of ML to current practices in health care (***Status quo is worse)***. Some of the harms of ML were recognized but were either believed to be justified by benefits (***Risks justify benefits***), minimized by being characterized as being irrelevant to the interviewee's work product (***Not in my AI***), or by downplaying the role of ML in health care, usually by locating ultimate decision-making authority with a clinician (**ML doesn't make decisions**). Similarly, other statements suggested that the harms of developers' products were not characteristics of the products themselves but arose from how they were used or misused (***Off-label use***). Finally, another type of statement stressed the role of regulation in assuring that harms would be minimized or prevented (***Regulation prevents harms***).

3.4. *Moral awareness and engagement*

In contrast, other participants made statements reflecting not only a recognition of the potential for their work to cause harm, but that their decisions had moral implications for which developers

had responsibility. There was almost no overlap between the set of participants who made statements indicating moral disengagement and those who made statements indicating moral engagement, which were defined as statements indicating awareness of a moral issue, statements recognizing the potential for conflicting interests leading to a moral dilemma, statements acknowledging responsibility, and statements indicating that responsibility aligned with personal values. Almost all statements indicating awareness of a moral issue also acknowledged some responsibility of ML developers, or at least recognized the role of developers in potentially causing harm.

> *You know, for some of these indications there are very negative effects to incorrectly identifying a person, either positively or negatively. Say the treatment for a certain indication puts somebody under a lot of duress and if we falsely flag somebody as having that indication then the culpability of that duress, you know, at least partly does lay on our shoulders.* (Participant 8)

Another participant demonstrated their awareness of a moral issue, as well as recognition of the link between design decisions and harms.

> *It's hard to realize that hey, somebody could actually not get treatment or a claim for somebody could be denied because you built a claims adjudicator algorithm, so that compass I think exists with us because you can fine tune your algorithm to be let's say more precise or be more specific, or like for precision recall, and both have different implications.* (Participant 24)

This participant went on to acknowledge not only the link between developers' algorithmic design decisions and effects on patients, but also the power and implied responsibility conferred by the data scientist's specialized knowledge.

> *Now, a data scientist has tremendous powers here because like your stakeholders don't really understand what precision recall is and where that threshold should be, so it's up to you to use your own judgment and say, you know what, actually I think I would rather that people have their claims paid than denied, so I will just tune it for true-positives.* (Participant 24)

A few of these participants also recognized moral differences of ML in the health care context:*...but there's a lot of consequences in telling people to do the wrong thing in healthcare.* (Participant 3) Others made statements reflecting a sense of responsibility for ensuring that their products would be of benefit to patients, and that fulfilling that responsibility was not only aspirational, but a requirement, and one that aligned with personal values.

> *But, you know, my hope is that also the people on the plans, like the members, will also benefit from these products. If I didn't think that they were going to be also benefiting from these products, then I probably wouldn't be working at [respondent's company].* (Participant 26)

Another participant also expressed that the purpose of their product was to benefit patients: *This is why we're here. This is why we're doing this is to help people. And I would like to think that we're helping people.* (Participant 25) But this participant also described a sense of internal

conflict about their goals: …*that's a fine line to walk every day, right, because at the end of the day like we're B2B products.* (Participant 25)

It was striking that many of the statements indicated recognition of the potential for conflicting interests leading to a moral dilemma, primarily financial interests: …*a lot of times you're just seeing predictive models being built for… based on cost, right… it's a very easy… easily understood outcomes, and then that leads to all sorts of potentially irrelevant or even slightly harmful socially or clinically sort of predictions being made.* (Participant 5) This participant indicated taking action to mitigate the potential harm: *… there was a separate analytics team … who did the predictive modeling work. But we were involved to help them determine some of the more useful inputs and also the outcome of interest, and we did steer them away from cost-based outcomes.*

However, several participants expressed discomfort with harms that could be inflicted by users of their products, and a lack of knowledge or ability to prevent those harms. *What are the safeguards we put in to make sure that when that genomic data gets other sources of data that it doesn't ever go near underwriters? You know, how do we quarantine that data that it's only used to improve patient outcomes…and never for estimating risk, you know, for the business side?* (Participant 1) Some even expressed resignation or inevitability of conflicting interests leading to misuse. …*and these are not things that I advocate nor does the entire… our company advocate at all, but…at the end of the day a company's gonna do what a company's gonna do.* (Participant 20)

4. Discussion

We conducted a qualitative analysis of interviews of 40 developers in the U.S. who were working on ML-based predictive analytics for health care. In our analysis of ML developers' perceptions of responsibility for harms of their work, we found that many of them raised issues indicating an awareness of a moral component of those harms – that is, that those harms could be caused by developers' actions (Figure 1: *Moral awareness)* and that developers or others might have responsibility to mitigate those harms. Few of these developers, however, described taking action to prevent or mitigate harms, possibly because of lack of knowledge about how to do so, or perceiving lack of agency (Figure 1: *Moral action*). However, developers also expressed uncertainty about responsibility for averting harms as an individual developer working in a company and moral conflict between personal values and those of their companies (Figure 1: *Conflict*).

One subset of developers, while recognizing harms, also displayed several forms of distancing themselves from harms or responsibility for those harms that were similar to a phenomenon described in the literature as moral disengagement. (Bandura, Barbaranelli, Caprara, & Pastorelli, 1996) Bandura *et al.* developed this construct as a cognitive mechanism to "deactivate moral self-regulatory processes and thereby help to explain why individuals often make unethical decisions without apparent guilt or self-censure." (Bandura, 1986) We do not claim that this cognitive mechanism is active in the ML developers that we interviewed, or make any claims about psychological processes in general. However, we do find similarities between the types of

rationales made by ML developers and researchers in other fields that serve to minimize harm, deflect responsibility for mitigating harm, or justifying research or its products despite the recognized harms. (White, Bandura, & Bero, 2009) Statements reflecting moral disengagement could be grouped into two general types: those that indicated minimization of the risks of ML (Figure 1: *Minimizing risk*), and those that indicated minimization of the ML developer's responsibility for those risks (Figure 1: *Minimizing responsibility*).

Our findings corroborate those of others who have found that AI developers have a number of rationales for their detachment from responsibility for their work. For example, in interviews of developers of health care AI developers, Vakkuri *et al.* heard several types of explanations that ethical concerns were not relevant to their work. One was that if projects were early-stage, i.e. "just a prototype," they didn't have any responsibility attached to them. (Vakkuri et al., 2020) Gotterbarn *et al.* (Gotterbarn, 2001) and McDonald *et al*. (McDonald & Pan, 2020) also found that computer scientists and students had a narrow view of responsibility that created moral distance by being task-oriented, by deflecting blame for errors (i.e. flaws in developers' programming being framed as "computer error"), or by casting failures in software as "inevitable or normal accident" inherent in complex systems. (Nissenbaum, 1994)

However, the subset of developers who not only recognized potential harms of their work, but also expressed a sense of responsibility for preventing or mitigating them was largely not overlapping with the group who made statements indicating moral disengagement. We do not know whether there were any particular characteristics that distinguish these two different groups of ML developers, such as education, experience with working in the health care context, role in the company, or demographic characteristics such as age, gender, race or ethnicity. We will investigate this question further in a larger sample of ML developers.

The financial conflicts of interest identified by our participants could be in part due to our sample being drawn almost completely from ML developers working at companies. That said, worries over how ML-based products might be misused in health care by health insurers and health care institutions were of concern to our interviewees. ML developers in corporate settings face not only internal values conflicts or uncertainty, but conflicts between their values and goals and those of their companies.

These findings suggest possible facilitators and barriers to the development of ethical ML that could act through encouragement of moral engagement or discouragement of moral disengagement. Regulatory approaches that depend on the ability of ML developers to recognize, accept, and act on responsibility for mitigating harms might have limited success without education and regulatory guidance for ML developers about the extent of their responsibilities and how to implement them, for example through standardization of key aspects of model evaluation such as performance metrics. Facilitators could include the integration of people with deep clinical knowledge on development teams, and alignment of organizational values with those of individual developers in order to reduce values conflicts, for example, about how to avoid misuse of MLPA models. Companies could also facilitate ethical ML development by encouraging a sense of agency among developers in making design decisions with values implications. However, the

conflicts of interest inherent in corporate settings and in MLPA products aimed at increasing health care efficiency pose particular challenges to mitigating their negative impacts. While our findings suggest internal actions that ML developers and companies can take to foster ethical ML developers, they also lend support to technology company arguments that regulation should come from government and not be developed themselves (Carter, 2020), and to those who question the ability of AI and data analytic companies to critically evaluate themselves. (Martin, 2022)

Table 1. Participants' Professional and Academic Characteristics

Participant Characteristics	**(n=40)**	**%**
Management levels*		
None	15	37.5%
Mid-level	9	22.5%
High-level	16	40.0%
Data interaction levels**		
Data only	15	37.5%
Data +	25	62.5%
Academic backgrounds		
Bachelors	11	27.5%
Health-related Masters	5	12.5%
Non-health-related Masters	6	15%
Health-related PhD	5	12.5%
Non-health-related PhD	9	22.5%
MD	4	10.0%
Type of organization		
Computer software and information technology - health care	29	72.5%
Computer software and information technology - general	3	7.5%
Health insurer	3	7.5%
Hospital	5	12.5%
Number of employees at organization		
å	19	47.5%
51-1,000	5	12.5%
Over 1,000	16	40.0%

*None refers to participants without managerial duties; Mid-level refers to participants with some managerial duties; High-level refers to participants with participants with extensive managerial duties
**Data only refers to participants who handle and work directly with the data in their daily work; Data + refers to participants who not only work with data but also perform other functions within their organization.

Table 2: Forms of moral disengagement identified in statements of MLPA developers

Moral disengagement type	Example
Minimizing risk	
No difference	*…it's like your financial data is out there too and somebody can way more ruin your life from, you know,*

Harms of ML are no different in health care than in other contexts	*stealing your identity than they can from like posting that so-and-so has… except for a couple of conditions, you know… like who cares what… like that's my attitude* (Participant 20)
Status quo is worse The status quo in healthcare (without ML) is worse than any hazards that ML might present	*So I mean we're expecting them to assimilate data, draw conclusions, and make projections, and when a computer does it somehow it seems more scary, but to me actually the fact that a person can just make a decision based on their gut is more scary…* (Participant 16)
Risks justified by benefits AI has risks but they are justified by benefits	*There's been all sorts of really terrible uses of machine learning that mostly penalize people that are already penalized in lots of other ways, like people of color or other kind of minorities. It's just sort of amplifying all these other bad things that are already happening….but I'm also not like a person… you know, I want to be able to do machine learning and have progress and see…machine learning helping medicine, 'cause it has so much that it can offer I think. (Participant 15)*
Not in my AI There may be hazards of AI, but they are not relevant to the type of AI that the participant works on	*I think that the problem of bias and pitfall might be more pertinent to other types of technologies, maybe like device technology. But I'm just… all my experience has been in the clinical decision support world where I really don't see a huge amount of risk.* (Participant 10)
Minimizing responsibility	
ML doesn't make decisions The healthcare provider makes the final decision, not the ML	*It totally leaves it in the clinician's hands. The clinician understands the context within which the prediction is made and they know that, you know, it's up to them to decide whether or not the patient should be treated. It's really just an indicator. It's like the dog in cartoons that points itself in an arrow, it says look this way, and so, you know, the clinician goes and has a look at the patient and they decide whether or not to treat them and how they should go about doing so.* (Participant 9)
Off-label use What other people do with produce is the problem, not the product itself	*I mean it depends on how the analytics is used and the purpose and the motives and the intention of the users. But as producers of analytics, we intend them to be used for general good I mean I would say.* (Participant 31)
Not my job I don't have the expertise or it's not my role	*I'm not like a health economist type of person, so my… the unsatisfactory answer is my work has not tried to optimize for any of that.* (Participant 17)

Regulation prevents harms Regulation is responsible for preventing harms	*… we raised that to the company and we talked about it and we sort of said okay, there's federal laws in place to prevent that from happening, so that's why, you know, we were sort of okay with that moving forward.* (Participant 26)

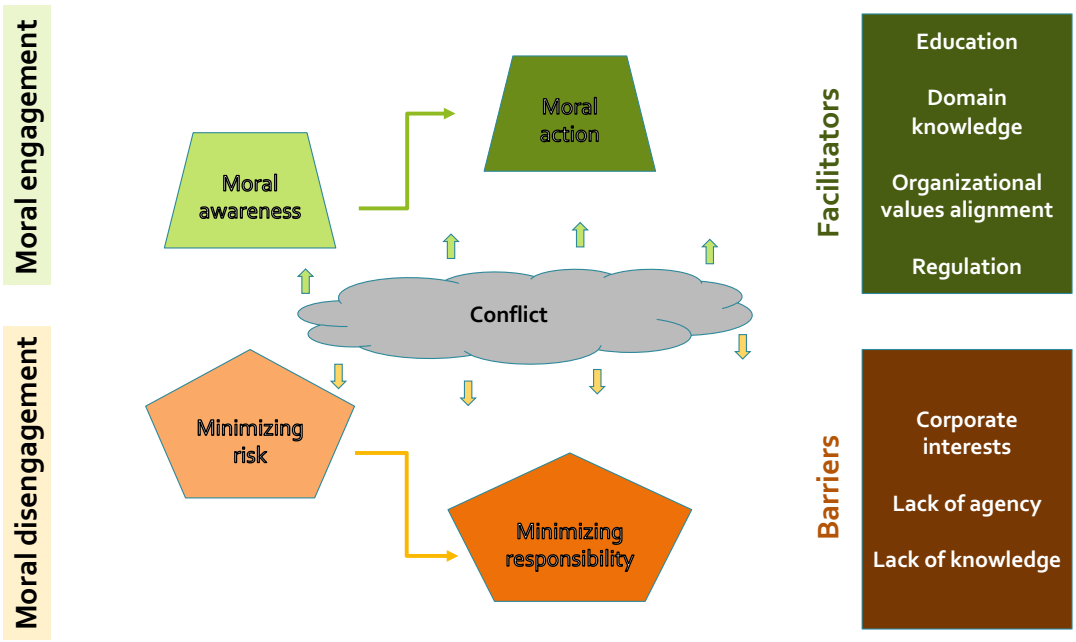

Figure 1: Facilitators and barriers to ethical ML

References

Bandura, A. (1986). *Social Foundations of Thought and Action: A Social Cognitive Theory*. Upper Saddle River, NJ: Prentice-Hall, Inc.

Bandura, A., Barbaranelli, C., Caprara, V., & Pastorelli, C. (1996). Mechanisms of Moral Disengagement in the Exercise of Moral Agency. *Journal of Personality and Social Psychology, 71*, 364-374.

Bernard, H. (2006). *Research methods in anthropology : qualitative and quantitative approaches*. New York: Altamira Press.

Carter, D. (2020). Regulation and ethics in artificial intelligence and machine learning technologies: Where are we now? Who is responsible? Can the information professional play a role? *Business Information Review, 37*, 60-68. doi:doi:10.1177/0266382120923962

Char, D., Shah, N., & Magnus, D. (2018). Implementing Machine Learning in Health Care — Addressing Ethical Challenges. *NEJM, 378*, 981–983. Retrieved from http://www.nejm.org/doi/10.1056/NEJMp1714229

Chen, I. Y., Pierson, E., Rose, S., Joshi, S., Ferryman, K., & Ghassemi, M. (2021). Ethical Machine Learning in Healthcare. *Annual Review of Biomedical Data Science, 4*(1), 123-144. doi:10.1146/annurev-biodatasci-092820-114757

Gotterbarn, D. (2001). Informatics and professional responsibility. *Sci Eng Ethics. , 7*, 221-230. doi:doi: 10.1007/s11948-001-0043-5

London, A. (2022). Artificial intelligence in medicine: Overcoming or recapitulating structural challenges to improving patient care? *Cell Reports Medicine, 3*, 2666-3791. doi:https://doi.org/10.1016/j.xcrm.2022.100622

Martin, K. (2022). *Ethics of Data and Analytics*: Taylor & Francis Group.

McDonald, N., & Pan, S. (2020). *Intersectional AI: A Study of How Information Science Students Think about Ethics and Their Impact.* Paper presented at the Proceedings of the ACM on Human-Computer Interaction.

Miles, M., Huberman, A., & Saldana, J. (2019). *Qualitative Data Analysis: A Methods Sourcebook* (4th ed.). Los Angeles: SAGE Publications.

Narvaez, D., & Rest, J. (1995). The four components of acting morally. Moral behavior and moral development: An introduction. . In L. Nucci, D. Narvaez, & T. Krettenauer (Eds.), *Handbook of moral and character education.* (pp. 385-400). Oxfordshire, UK: Routledge.

Nichol, A. (2022). *Facilitators and barriers to ethical machine learning in healthcare: A qualitative study on developer perspectives of potential harms.* Paper presented at the The 5th ELSI Congress, Virtual. https://elsicon2022.us2.pathable.com/people/Jmvd8qeFP3DKDbMph

Nichol , A., Batten, J., Halley, M., Axelrod, J., Sankar, P., & Cho, M. (2021). A Typology of Existing Machine Learning–Based Predictive Analytic Tools Focused on Reducing Costs and Improving Quality in Health Care: Systematic Search and Content Analysis. *JMIR, 23*. doi:DOI: 10.2196/26391

Nissenbaum, H. (1994). *Computing and accountability.* Paper presented at the Communications of the ACM.

Strauss, A., & Corbin, J. (1997). *Grounded Theory in Practice*. Los Angeles: SAGE Publications.

US Food & Drug Administration. (2018). *Developing a Precertification Program: A Working Model*. Retrieved from https://www.fda.gov/media/112680/download

US Food & Drug Administration. (2019). Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) - Discussion Paper and Request for Feedback. Retrieved from https://www.fda.gov/media/122535/download%0Ao

US Food & Drug Administration. (2021). *Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) Action Plan*. Retrieved from https://www.fda.gov/media/145022/download

Vakkuri, V., Kemell, K., Jantunen, M., & Abrahamsson, P. (2020). *"This is Just a Prototype": How Ethics Are Ignored in Software Startup-Like Environments.* Paper presented at the Agile Processes in Software Engineering and Extreme Programming: 21st International Conference on Agile Software Development, Copenhagen, Denmark.

White, J., Bandura, A., & Bero, L. (2009). Moral disengagement in the corporate world. *Account Res., 16*, 41-74. doi:doi: 10.1080/08989620802689847.