Opening the Black Box: Educational Machine Learning Videos for a General Public Audience

Harini Suresh
hsuresh@mit.edu
Massachusetts Institute of Technology

Natalie Lao
natalie@mit.edu
Massachusetts Institute of Technology

ABSTRACT

While machine learning (ML) affects increasingly more and increasingly consequential aspects of life, the general public does not have a good understanding of how ML works or what it can and cannot do. As attention shifts to regulating and legislating these technologies, the public should be empowered to engage in related discussions. Understanding begins with education, but many popular and currently available materials for learning ML are too technical to be accessible to a general audience, too broad to be useful, or simply wrong. We describe the process of piloting short educational videos with the goal of empowering the public to think and debate critically about the impact ML can have on society. An important facet of our methodology is to integrate ethical and societal considerations into each technical video topic. Teaching responsible ML throughout the entire educational pipeline has the potential to create an environment where people can critically discuss ML and demand more accountability from the companies producing ML products. In this paper, we explain the content decisions made to best serve this audience. In addition, we demonstrate improvements in ML understanding and attitudes towards civic engagement using a small-scale study with participants of different knowledge levels.

CCS CONCEPTS

• Computing methodologies → Machine learning; **• Social and professional topics → Computing education**; **• Human-centered computing → Human computer interaction (HCI)**.

KEYWORDS

machine learning education, ethical machine learning, educational videos

1 MOTIVATION

Despite a rapid expansion of machine learning (ML) across fields and industries, it is not seen as understandable by the general populous. A 2017 study by The Royal Society interviewed members of the public in the UK, finding that a majority of participants knew "little to nothing" about machine learning [18]. While many of the participants were aware of technologies that use ML, very few were aware of how the technology worked, "even at a broad conceptual level." Another study found that even amongst UX designers working on projects that involved ML, a lack of understanding was common. One participant referred to ML as "black magic," stating that "designers don't understand the constraints of the technology and how to employ it appropriately"[9].

As ML technologies appear in more everyday contexts and make increasingly consequential decisions in our lives, this lack of understanding is troubling. Regulation around these technologies is nascent, and as policymakers think about reasonable legal structures, the public should be empowered to engage in these discussions.

Widespread public engagement can also help disrupt a harmful power dynamic between the producers of automated systems and the people that these systems affect. The companies building ML technologies consist of a mostly homogeneous population that are overwhelmingly wealthy, white, and male. In 2018, only 18% of first authors at 21 ML conferences and 15% of Facebook's AI Research staff were women [6], and only 2.5% of Google's workforce was black [7]. A recent report by the AI Now Institute details the way in which this lack of workplace diversity is fundamentally tied to gender- and race-based discrimination in systems themselves [22]. The public's general lack of understanding of how these systems work only exacerbates the power differential. If individuals and communities are able to critically understand the impact and limitations of ML, they can hold developers and companies accountable, and perhaps feel empowered to build grassroots technologies themselves.

In this paper, we describe piloting an educational YouTube channel aimed at improving public knowledge of ML, including an initial evaluation of efficacy. Our videos are intended to be understandable to a high school and above audience, are 5-10 minutes long, and feature animated hand-drawn illustrations. The goal of this effort is to empower people to (1) develop a balanced understanding of the potential benefits and risks of ML technologies, (2) engage in educated civic discussions about these technologies and (3) recognize the ML technologies in products and systems and demand accountability from their producers. Eventually, we hope that some of these viewers will go on to create new, responsible technologies themselves.

In Section 2, we discuss related efforts in ML education. In Section 3, we present our content decisions and walk through an example video, and in Section 4 and 5, we demonstrate the impact of a particular video on viewers by analyzing the results of a study with pre- and post- surveys.

2 RELATED WORK

There are a wealth of educational materials available about machine learning, but few that are easily understandable and useful to the majority of adults who are not currently pursuing an education or career in ML, data science, or computer science-related fields.

2.1 Courses

Full-semester undergraduate or graduate courses on ML are available to enrolled students at many educational institutions as well as on online platforms (e.g., MIT Open Courseware [5] or Stanford

Online [8]). These courses are typically intended for computer science (CS) students with an existing background in the material, and aim to provide a significant level of technical depth as well as a wide breadth of ML topics.

Shorter, primarily online courses or materials are also offered by some companies (e.g., Coursera [19], fast.ai [3], or Google [2]). These courses may be more targeted (for example, towards engineers focused on practical implementation) or may cover less depth than university courses. Some recent efforts, such as Embedded EthiCS @ Harvard [13] or the Responsible Computer Science Challenge [4], have produced materials focused on ethics that are intended to integrate into undergraduate CS courses.

In general, courses require more commitment from the audience (usually lasting from several hours to entire months). They usually build upon material covered in previous units and are not meant to be broken up and consumed as single lessons. However, many adults who wish to know something about ML will not have the ability or willingness to invest significant time into an entire course. Even of those who start open online courses, there tends to be a steep drop-off in viewers after one to two weeks [10, 21]. Additionally, most full courses on ML are designed for audiences with at least some technical background. Instead, our focus is on the general adult population learning in an informal and self-motivated way, prompting significantly different content and platform decisions.

2.2 Singular Materials for Self-Learning

Outside of full courses, one-off materials such as blog posts, YouTube videos, Quora answers or news/media articles are a common source of self-education on ML. These materials typically discuss a specific topic, and range in technical depth, length, and structure. For adults not enrolled in an educational institution, online materials like these are an easily accessible, low-commitment source for self-learning.

However, unlike full courses, many of those one-off materials are not created by experts or professionals, and are not vetted or checked for errors. As a result, many of the resources currently available online fall short in different ways, including:

(1) **They are intended for a technical audience, and therefore are confusing and/or intimidating to people without significant prior knowledge.** For example, several articles on the first page of Google search results for "what is a neural network?" immediately present the reader with network diagrams and equations [20, 23]. While this information might be useful to people with technical experience, it is likely overwhelming or confusing to viewers from other backgrounds.

(2) **They don't cover enough depth, thus misrepresenting problems and solutions in ML.** For example, a 6-minute YouTube video "What is Artificial Intelligence (or Machine Learning)?" with approximately 600,000 views describes AI and ML as "not only programming a computer to drive a car by obeying traffic signals, but it's when that program also learns to exhibit signs of human-like road rage" [16]. Oversimplification and personification of ML mislead beginning learners and misrepresent the human element of creating such products.

(3) **They are factually incorrect, or provide hand-wavy and unclear explanations.** For example, many resources describe neural networks as imitating the form and function of the brain's optical nerves or primary visual cortex. While neuroscientists are working to create machine learning models that mimic the human brain, current ML technologies do not reflect human biology.

By creating videos that are short (5-10 minutes long) and can stand on their own, we aim to capture the low commitment of one-off learning opportunities while avoiding the pitfalls outlined above. At the same time, we also share content that covers a wide breadth of ML topics in enough depth to be useful, providing viewers with the opportunity for comprehensiveness as in a full course.

3 OUR CONTENT

We aim to improve public knowledge of ML by creating short, engaging online educational videos to teach ML such that people are empowered to (1) think and debate critically about the impact ML can have on society and (2) eventually pursue their own projects. Moreover, we believe that teaching responsible ML throughout the entire educational pipeline will create an environment where consumers and developers instinctively demand more accountability for the societal impacts of ML systems and products. The goal of our content is help a general audience develop into consumers who are able to critically and conscientiously discuss the implications of ML in society, and then guide some of those consumers in becoming ethical creators of ML applications themselves.

3.1 Videos as an Effective Content Modality

The average American spends less than 5% of their life in traditional classrooms, and research increasingly shows that most scientific knowledge is acquired outside of school [11]. As a result, informal learning resources have been described as a "cost-effective way to significantly improve public understanding of science."

We choose to use videos to communicate ML topics rather than articles, books, or other text-based modalities that are commonly used. Videos as instructional tools have been shown to significantly improve recall of conceptual information and creative problem solving over other mediums such as text or labeled images, particular for participants with less prior knowledge [17].

YouTube in particular has grown to be an internationally popular platform for general education. Google has announced that they will be increasing funding and support for "EduTubers" specifically, with a $20 million grant in late 2018 [1].

3.2 Content Decisions

3.2.1 Prioritizing standalone videos. To avoid a prohibitive time commitment, we wanted to ensure that most individuals with a high school level education could pick any individual video and understand it. In addition, we prioritize making short videos; research has indicated that educational videos longer than six minutes result in significant viewer drop-off [12, 14]. This allows us to reach a much broader audience than courses, which typically require participants go through all prior material in order to move on to the next lesson. Moreover, this audience who may not have the time or interest to commit to learning about ML is precisely who we wish to target: adults who do not work in or study ML, but who nonetheless are affected by it daily. While our videos may still build on each other and watching several or all of them might lead to a deeper, more comprehensive understanding of the subject, this is not be a requirement.

As a result, in our videos we motivate various ML topics in a heavily example-based way, rather than by building on theory or other methods. This aids in making the topics immediately understandable and relatable on their own.

3.2.2 Content Categories. With the goal of empowering the public to participate in educated discussions about ML technologies, we identified several categories of content to provide:

(1) ML Basics: These videos are focused on building blocks of ML. These include important concepts (e.g., false positive/negative rates, tensors/matrices) and some review of more advanced high school math concepts (e.g., derivatives, probability). Each of these videos is around five minutes, and explains the topics as they are relevant to ML. We expect that most viewers will already have some knowledge on these topics, but we provide these videos so people who are not familiar can still follow later videos.

(2) Explaining ML: These videos make up the bulk of the content. They cover ML methods (e.g., neural networks), applications (e.g., photo filters), interpretability methods (e.g., bounding boxes), and high-level overviews (e.g., steps of a typical ML product). The videos range in length from around 5-10 minutes, and walk through specific example to ground the explanations. We bring considerations of societal impact, risks and limitations into each video in an effort to give viewers a nuanced and balanced knowledge of ML.

(3) Debating ML: These videos delve into other topics that are not necessarily focused on introducing new material, but on exploring and discussing relevant topics. They include videos on relevant policies or current events (e.g., face recognition ban), discussions on potential ethical concerns with ML (e.g., risks ML-based police body cameras), and question-and-answer style discussions with guests.

3.2.3 Presentation Style. The style of our videos is informal, easy to understand, and short, in keeping with our goal of making our content accessible and friendly. The majority of each video features chalkboard-style drawings and animations with a voiceover. In general, animations have been shown to improve viewer engagement and learning outcomes in instructional videos when compared to static images [15]. There are also sections featuring presenters speaking directly to the camera, sometimes with text or animations overlaid.

3.3 Video Example: Stages of ML

In this section, we describe the content of an example video in more detail, including static screenshots of the animations.

This particular video walks through the process of creating an ML product, from problem definition to post-deployment. We chose this topic as one of our first videos to give viewers important context and a basis for considering the societal impacts and limitations ML from the start. This video demonstrates the idea of integrating ethical considerations directly into teaching ML content. For example, rather than make a separate video about analyzing an ML model’s performance across subgroups, we introduce this concept directly when describing the ‘evaluation’ stage of an ML product, framing it as a clear concern that should instinctively arise.

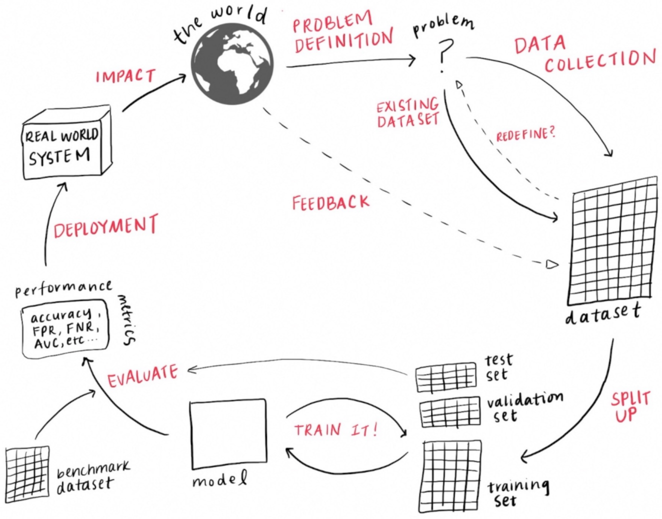


Figure 1: A screen from the video describing stages of ML, depicting the entire process. In the video, each stage is discussed in detail.

Throughout the video, each step is added to a cyclic diagram. We choose this representation to emphasize the fact that when ML is implemented in the real life, populations shift and a product’s impact may take time to manifest, warranting continuous re-evaluation. We introduce and use the running example of building a mobile app to classify whether a friend is sad or not. A screenshot of the finished diagram can be seen in Figure 1, but we describe how we broke down the steps in more detail:

(1) **Problem Definition.** We chose to include this as an important step in the ML pipeline. Arriving at a problem statement is process that introduces specific limitations and assumptions. However, this step is often skipped over in traditional ML courses, where assignments typically come with a predefined problem to solve but without a discussion of why that particular framing was chosen, its implications, and its tradeoffs. In the context of detecting whether or not a friend is sad from an image, we discuss how the entire premise of the problem statement rests on the assumption that sadness can be detected from an image, and how deciding on a binary classification framing assumes that most people exhibit a single dominant emotion.

(2) **Data Collection Step 1: Choosing a Population.** We spend a significant fraction of the video describing the process and consequences of data collection. This also tends to be skimmed over or not mentioned in traditional ML courses, where participants are usually given a dataset without much time spent discussing where it came from and its limitations. We first describe the process of defining a population (Figure

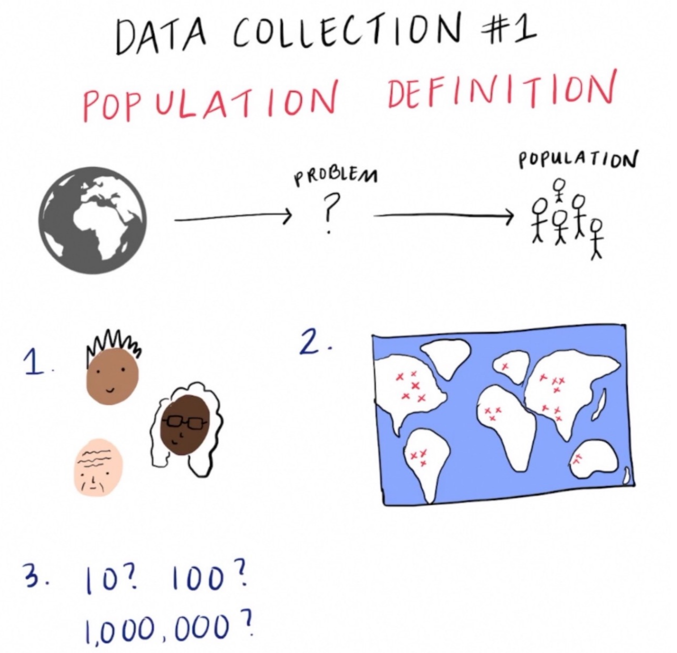


Figure 2: A screen from the video describing stages of ML, depicting the process of choosing a population during data collection.

2), raising several questions about who the data represents and how similar or varied this demographic is. Here, we also introduce the concept of a model's dependence on data by describing how a non-representative dataset can lead to poor performance for other parts of the population.

(3) **Data Collection Step 2: Measuring Features from the Population.** In this step, we describe the process of going from an actual population to a dataset of measurements about this population. We bring up questions to ask about this step, such as whether the quality of images is consistent across the population, and what other information should be collected and included. This step also describes the process of collecting labels. Crucially, this is described as a collection *process*, and is not abstracted away as existing ground truth. In the running mobile app example, we describe possible methods of label generation such as gathering self-reported labels from participants or having other people hand-label them, and how any method will introduce a specific type of noise into the data that carries consequences.

(4) **Data Collection Alternative: Existing Sources.** We describe the wealth of pre-existing datasets that are available online, and how many ML products simply leverage existing data rather than going through steps (2) and (3). We describe how this often involves massaging the pre-existing dataset into a usable form, introducing new sources of noise and possible bias. We emphasize that even pre-existing data is the product of a process and that all the questions and potential concerns brought up in the data collection description are still relevant and important.

(5) **Splitting the Data.** This step introduces the concepts of training, validation, and testing datasets, particularly emphasizing the importance of setting aside testing data during the development process.

(6) **Defining a Model.** This step describes how different types of data (e.g., timeseries vs. text vs. images) may require different types of models that are particularly suited to that type of data. Though we do not go into technical depth in this overview video, we mention that the purpose of a model is to pick up on patterns in the data, and that more complex data requires larger, more complex models (and vice versa).

(7) **Training the Model.** In this step, we explicitly introduce the idea that models simply learn from data with an optimization procedure. Here, the role of validation data in model selection is also mentioned.

(8) **Evaluating the Model.** This step goes through several important concerns that arise in model building. We discuss the fact that evaluation metrics are calculated on the test data or other benchmark datasets that carry the same concerns about quality and representativeness as brought up in steps (2) and (3). We bring up the choice of evaluation metric(s), and how different metrics may have different consequences for specific parts of the population; this is contextualized with the potential risks of false positives versus negatives in the mobile app example. Here, the importance of subgroup evaluation is also emphasized.

(9) **Deployment Preparation.** This step discusses practical implementation questions when trying to ensure that a model is actually used correctly and effectively. This is also a step that is typically out-of-scope of traditional ML courses. We describe the importance of visualization and interfacing for end users to correctly understand the model's predictions, and the need for some way to incorporate feedback about incorrect predictions.

(10) **Deployment and Scaling Up.** Here, we describe additional concerns that should arise when deploying a product or system, such as population drift, data privacy and consent.

(11) **Impact.** The final section emphasizes the importance of continually monitoring an ML system once it is deployed to ensure that it is actually having the expected impact, and the need to go back and re-evaluate or update various steps if it is not.

4 VIDEO EVALUATION METHODOLOGY

In order to gain an understanding of the efficacy of our videos at teaching machine learning concepts to a broad audience, we ran a study to analyze how and if audiences of various ML knowledge backgrounds retained the information provided through our *Stages of ML* video.

4.1 Study Design

The study consisted of an electronic pre- survey and post- survey via Qualtrics. Both the pre- and post- survey were designed to take less than 5 minutes to complete, excluding the time it took to watch the 10-minute video.

The pre- survey consisted of three parts: (1) Demographics questions, which also included items that gauged the participant's background knowledge of ML; (2) ML concept understanding and perception questions; and (3) the *Stages of ML* video followed by a question of whether or not the participant had seen the video before. Participant emails were collected in order to administer the post- survey with a 2-day time delay to measure knowledge retention and avoid short-term learning effects on the data. The post-survey consisted of only the questions from section (2) of the pre-survey. We designed the survey questions on ML knowledge and perceptions ourselves because there are few to none prior validated instruments for this purpose in the ML education research space.

There were 8 demographics questions. The first 5 questions were general questions on the participant's gender, ethnicity, age range, and education. The last 3 questions helped gauge the participants' prior knowledge of ML, and are presented below:

- D6: Which of the following types of courses, if any, have you taken in person (i.e. at a high school or university)? *Type: Multiple answer, multiple choice.*
  - Machine Learning
  - Artificial Intelligence
  - Data Science
  - Probability or Statistics
  - Calculus
  - Linear Algebra
  - Any other Computer Science course not listed above
  - None of the above (*Exclusive choice.*)
- D7: Which of the following types of courses, if any, have you taken online? *Type: Multiple answer, multiple choice. Same choices as above*
- D8: How would you consider your own general knowledge of machine learning? *Type: Single answer, multiple choice.*
  - I don't know anything about machine learning
  - I've heard of machine learning in passing
  - I've read media articles or listened to news about machine learning technologies
  - I've used machine learning/AI-based tools for work or at home
  - I've read technical journals/research papers on machine learning

There were 8 questions measuring ML concept understanding and perception, with 2 free response questions about understanding and 6 Likert scale questions about perception as presented below:

- C1: How would you explain machine learning to a friend who had never heard of it before? (Just try your best. If you don't know what to say, you can write "I'm not sure.") *Type: Free Response.*
- C2: Can you list three common things you see or use everyday that use machine learning? *Type: Free Response. Numbered lines provided.*
- P1: Please rate how much you agree with the statement: I feel confident explaining or discussing machine learning with a non-technical person. *Type: 5 point Likert scale (Strongly Disagree, Disagree, Neutral, Agree, Strongly Agree).*
- P2: Please rate how much you agree with the statement: I feel confident explaining or discussing machine learning with a machine learning expert. *Type: 5 point Likert scale.*
- P3: Please rate how much you agree with the statement: I feel invested in future machine learning legislation or policy decisions. *Type: 5 point Likert scale.*
- P4: Please rate how much you agree with the statement: I feel confident voicing my opinions relating to machine learning legislation or policy decisions. *Type: 5 point Likert scale.*
- P5: Please rate how much you agree with the statement: Machine learning will make everyone's life better. *Type: 5 point Likert scale.*
- P6: Please rate how much you agree with the statement: Machine learning is dangerous. *Type: 5 point Likert scale.*

The post- survey was administered at least 48 hours after the participant submitted the pre- survey and watched the video. Access to the post- survey was granted through email and was open for a maximum of one week.

4.2 Recruitment

In order to reach a broad audience with varying levels of ML background knowledge, we recruited volunteers through web forums focused on survey-taking, general purpose email lists, and CS/ML email lists at a local university. Participants who completed both parts of the study were entered into a raffle for five $20 gift cards.

4.3 Participant Demographics

There were 47 participants who completed both parts of the study. These participants were grouped into low knowledge (LK), medium knowledge (MK), and high knowledge (HK) in order to help us understand how our educational content and style affect these groups differently. The *Stages of ML* video is targeted towards a general audience and gives an overview of the machine learning process instead of deep diving into specific technical methods, so we expected to see higher increases in measures of ML concept understanding in lower prior knowledge participants. Additionally, the *Stages of ML* video explains ethical considerations and problems of fairness and bias throughout the entire pipeline in a way that traditional ML education resources do not provide, so we hoped to see some impact in measures of ML perceptions across participants of all prior knowledge levels. Participants were sorted into the three prior knowledge groups based on their pre- survey responses to D6, D7, and D8.

Participants who selected *I don't know anything about machine learning* as a response to D8 were sorted into LK. Participants who selected *I've heard of machine learning in passing* or *I've read media articles or listened to news about machine learning technologies* as a response to D8 were sorted into either LK or MK. For those participants, if they had taken Machine Learning or Artificial Intelligence as responses to D6 or D7, or if they had taken 3 or more other ML-related courses, they were sorted into MK. The rest of those participants were sorted into LK with two exceptions based on the researchers' observations that their 3 responses to C2 of the pre-survey were correct and too specific for them to belong in the LK group (C2 was not used in our analysis). Participants who selected *I've read technical journals/research papers on machine learning* as a response to D8 were sorted into HK. Participants who selected *I've used machine learning/AI-based tools for work or at home* as a response to D8 were sorted into either MK or HK. For those participants, if they had taken Machine Learning or Artificial Intelligence as responses to D6 or D7, they were sorted into HK. The rest of those participants were sorted into MK.

There were 11 people in the LK group, 19 in the MK group, and 17 in the HK group. The genders and ethnicities of the participants were generally similar across groups (in each group, the majority of participants identified as female and White). LK was 91% female and 9% male; MK was 74% female and 26% male; and HK was 71% female, 24% male, and 6% gender variant/non-conforming. LK was 73% White, 18% Asian, and 9% Black or African American; MK was 53% White, 37% Asian, 5% Hispanic, and 5% Prefer not to say; HK was 53% White, 41% Asian, and 6% Black or African American.

The distributions of age range and types of advanced degrees varied more widely across the three groups. The LK group had a near-normal distribution of age ranges from 18-24 to 65-74, peaking at 45-54 with 27%. The MK group was heavily skewed right, with 42% of participants aged 18-24 and a long tail going to 55-64. The HK group was more homogeneous, consisting of 82% in the 18-24 range and 18% in the 25-34 range. For the LK group, Non-STEM degrees were the top type of advanced degree obtained. For the MK and HK groups, other STEM degrees aside from math and CS was the top type of advanced degree obtained. The comparisons of age and degrees attained across the groups can be seen in Figure 4.

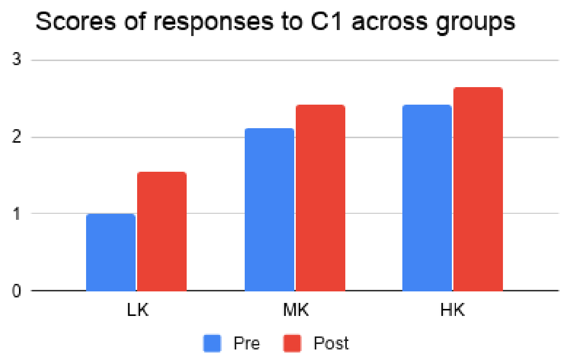


Figure 3: Scores for C1: "How would you explain machine learning to a friend who had never heard of it before?" for each knowledge group across pre- and post- surveys (rated at levels 1-3, where 1 is lower).

5 RESULTS

5.1 Improved Explanations of ML

The pre- and post- survey responses to question C1 (*How would you explain machine learning to a friend who had never heard of it before?*) were analyzed in two stages.

First, the responses were graded on a scale of one to three. A score of one indicated very little or mistaken understanding (e.g., "I'm not sure" or "learning from a computer"), a score of two indicated partial or vague understanding (e.g., "A lot of it is about gathering data and writing programs to work with that data."), and a score of three indicated correct understanding (e.g., "Using computers to look for patterns in a large dataset, which can then be applied to other situations to solve problems, make predictions, identify something, etc."). The results of this analysis are displayed in Figure 3.

Themes	LK	LK	MK	MK	HK	HK
	Pre	Post	Pre	Post	Pre	Post
data as an input	9%	**64%**	53%	**79%**	65%	**88%**
learning from data	-	**55%**	53%	**79%**	65%	**82%**
statistical processes	-	-	**11%**	-	**18%**	12%
testing the model	-	-	5%	5%	-	**6%**
task/outcome-oriented	**9%**	-	37%	**63%**	76%	**94%**
cyclic process	-	-	**11%**	5%	-	**6%**
societal impact	-	**9%**	-	**5%**	-	-

Table 1: Percent of answers with a given theme for question C1: "How would you explain machine learning to a friend who had never heard of it before?" for each knowledge group across pre- and post- surveys.

As expected, both the pre- and post- answer scores increase across knowledge groups. Within each group, there is an increase in scores in the post- survey as compared to the pre- survey answers. This increase is more pronounced for LK (0.55) than for MK (0.31) and HK (0.24). In other words, the marginal increase in answer correctness was largest for the lowest knowledge group, indicating highest efficacy for our video's target audience. An example of a pre- and post- survey answer improvement for an LK participant is:

- Pre: "I'm not sure" (Score: 1)
- Post: "Machine learning is analyzing data to make more accurate descriptions of societal processes." (Score: 2)

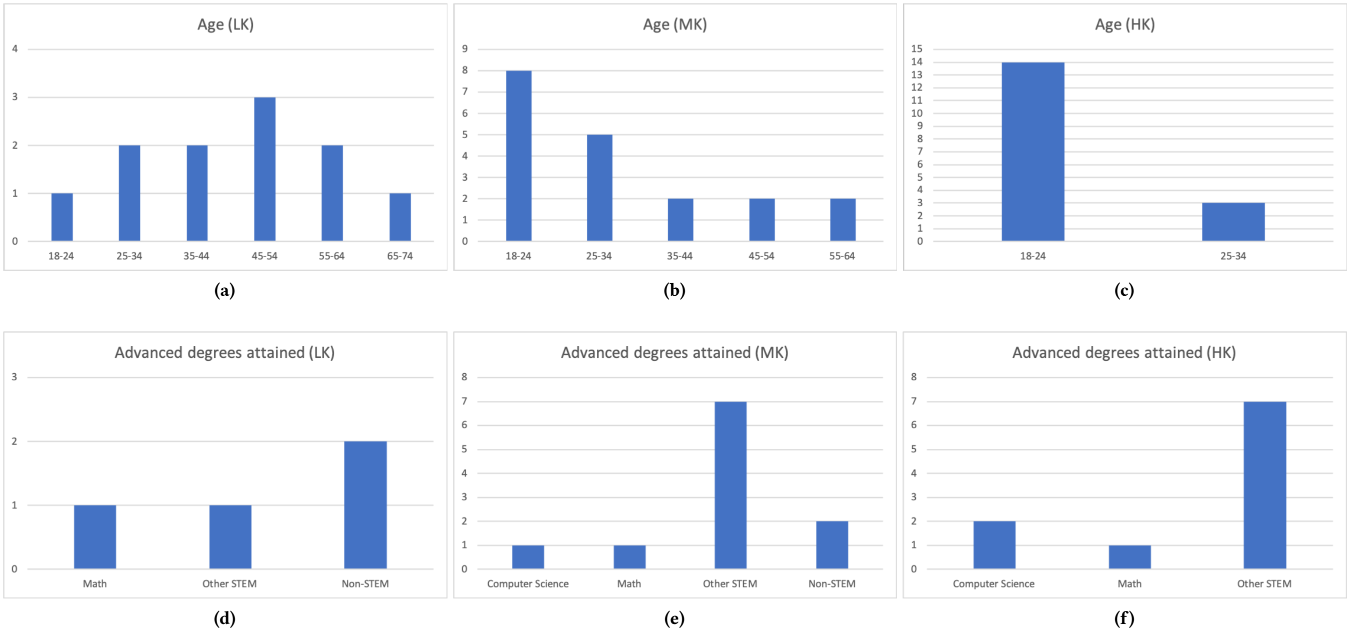
And an example of a pre- and post- survey answer improvement for an MK participant is:

- Pre: "Using mathematical methods to guess at and then validate the best model for predicting something desired." (Score: 2)
- Post: "Using data to train a computer model such that it will be able to make accurate predictions on new but similar data." (Score: 3)

Second, prominent themes were identified and tallied across answers. These themes, and their prevalence in pre- and post- survey answers for all groups, is displayed in Table 1. In the pre-survey, the lowest knowledge group's answers contained some incorrect themes (these are not displayed in the table), such as "humans learning using computers" or "making machines better," which were no longer present in any answers in the post-survey.

Of particular note is the increase in the prevalence of the "data as an input" and "learning from data" themes across all knowledge groups. This increase in prevalence is largest for the LK group, where only one participant mentioned data in the pre-survey, but the majority of participants did in the post-survey.

MK and HK groups exhibited a 26% and 18% increase in the "task/outcome-oriented" theme, which did not come out in the



(a) (b) (c)

(d) (e) (f)

Figure 4: Subfigures (a) – (c) display the age range distributions of the participants across the low, medium, and high knowledge level groups. (d) – (f) display the distributions in types of advanced degrees attained for the participants across the three groups.

LK group, indicating that different knowledge groups might gain different levels of insights from the video. There was also a slight increase in the "societal impact" theme. An example of an MK answer pair which includes this theme in the post-survey answer is:

- Pre: "Using computer algorithms to do stuff that we humans don't want to do, as it's boring. You train it using a data set of information, and the computer 'learns' based off that information. Then you set it loose to do it's thing, and review/correct it when needed." (Score: 3)
- Post: "So, you have a computer system that will do the boring stuff that us humans don't want to do. But in order to train it, you need a really good data set, otherwise you will input bias into it's thinking/training. Then it will not learn correctly." (Score: 3)

5.2 Changing Perceptions of ML

The pre- and post- survey responses to Likert scale questions P1-P6 for all three groups are presented in Table 2. For these items, Strongly Disagree is coded as 1 and Strongly Agree is coded as 5. In our analysis, we considered post-/pre- differences of 0.47 or above, as it reflects a shift of approximately half of a point within the Likert scale (10% of the total range). These differences were observed for 5 out of the 6 questions in the LK group, 1 out of 6 in the MK group, and 1 out of 6 in the HK group.

For P1 – P4, which deal with participants' confidence and feelings of responsibility about ML technologies' role in society, we unsurprisingly see that the base values for LK, MK, and HK show an ascending pattern. We also observe an increase from pre- to post- for these 4 questions across all knowledge levels.

P1 and P2 measured how confident the participant felt discussing ML with either a non-technical person or an expert. The LK group's P1 was the item for which there was the largest difference in pre- and post- responses. LK participants reported an increase of 0.91 (nearly an entire point) in confidence regarding explaining or discussing ML with a non-technical person. LK participants also experienced an increase in 0.64 (over half a point) in confidence regarding explaining or discussing ML with a ML expert. MK participants also experienced an increase in 0.53 (over half a point) in confidence regarding explaining or discussing ML with an ML expert. This indicates that the video was able to help people with less knowledge about ML feel more comfortable engaging in discussions about the topic.

P3 and P4 captured participants' general feelings about the importance of ML legislation/policies, and whether or not they would feel comfortable participating in forming such rules. The LK group once again saw an increase in over half a point (0.55) for both P3 and P4, implying that the instrument helped a lower prior knowledge audience feel more empowered and engaged in the societal/legal discussions surrounding ML. The HK group also saw an increase in nearly half a point (0.47) for P3, with the post- survey responses averaging at 3.94 (agreeing that they "feel invested in future machine learning legislation or policy decisions").

P5 and P6 measured positive and negative feelings towards ML technologies in general. While none of the three groups' responses to P5 (positive feelings towards ML) seemed to be significantly affected by the video, it is interesting to note that the LK group saw

Question	LK Pre	LK Post	Diff	MK Pre	MK Post	Diff	HK Pre	HK Post	Diff
P1: I feel confident explaining or discussing machine learning with a non-technical person.	1.45	2.36	**0.91**	2.89	3.26	0.37	3.94	4.24	0.29
P2: I feel confident explaining or discussing machine learning with a machine learning expert.	1.27	1.91	**0.64**	1.74	2.26	**0.53**	2.71	3	0.29
P3: I feel invested in future machine learning legislation or policy decisions.	2.27	2.82	**0.55**	3.16	3.26	0.11	3.47	3.94	**0.47**
P4: I feel confident voicing my opinions relating to machine learning legislation or policy decisions.	1.82	2.36	**0.55**	2.68	3	0.32	2.94	3.18	0.24
P5: Machine learning will make everyone's life better.	3.09	3.18	0.09	3.47	3.53	0.05	3.53	3.47	-0.06
P6: Machine learning is dangerous.	2	2.55	**0.55**	3.05	2.79	-0.26	2.71	2.71	0

Table 2: Average responses to 5 point Likert scale questions on perceptions of ML for each knowledge group across pre- and post- surveys. (Strongly Disagree = 1, Strongly Agree = 5).

an increase in over half a point (0.55) for P6, which captures fear of ML.

6 DISCUSSION

Our results suggest that our content and presentation methodology can be effective for improving ML knowledge as well as shifting perceptions about civic participation in issues relating to ML, particularly for groups with minimal prior knowledge. The increased awareness of important themes such as models learning from data is promising, as is the increased confidence discussing and voicing opinions about ML legislation. A participant in the 55-64 age range commented that: "Being an older learner and one steeped in books and printed materials, I always find educational videos to go too fast...your video was not as fast as many out there – I felt I could follow your arguments along the way. That diagram is wonderful!" Such feedback shows that our video content and style has promise at successfully reaching much broader audiences that are not included in the audience pool for traditional ML education resources. We also posit that our educational framework of teaching ethical concerns in each step of the ML pipeline helps better inform people about where ML can go wrong, which helps them better understand, retain, and articulate the potential problems with ML.

The results presented here are inherently limited in their size and diversity. In order to arrive at more conclusive findings and to better motivate future content, a more extensive recruitment procedure should be undertaken. Additionally, while we chose a video broad in scope and characteristic of the type of content we provide, it is only one video. A more comprehensive study with different types of videos could better gauge the efficacy of our content on the whole.

7 CONCLUSION

ML will affect our lives in increasingly more consequential ways, but the public is currently not well-equipped to engage in these discussions. Most educational efforts to this end focus on audiences with a significant prior technical background, but this makes up a small fraction of the population that does not reflect the general public. Through short and engaging educational videos, we aim to empower the public to understand and engage in critical discussions around these technologies. An understanding of the potential benefits, risks and limitations of ML can help consumers hold companies accountable and disrupt the harmful power dynamic between the companies that produce ML and the people it affects.

In this paper, our contributions are:

1. Motivation for video-based ML education for a general public audience, integrating limitations and societal considerations into each facet of ML.
2. A justification of content decisions made to best serve this audience.
3. An evaluation of the effect of a characteristic video on participants with different levels of prior knowledge, demonstrating an improvement in ML understanding and an increased willingness to participate in civic discussions about ML.

REFERENCES

[1] 2018. YouTube Learning: Investing in Educational Creators, Resources, and Tools for EduTubers. youtube.googleblog.com/2018/10/youtube-learning-investing-in.html

[2] 2019. Machine Learning Crash Course. https://developers.google.com/machine-learning/crash-course/

[3] 2019. Practical Deep Learning for Coders, v3. https://course.fast.ai/

[4] 2019. Responsible Computer Science Challenge. https://foundation.mozilla.org/en/initiatives/responsible-cs/

[5] Hal Abelson. 2008. The creation of OpenCourseWare at MIT. *Journal of Science Education and Technology* 17, 2 (2008), 164–174.

[6] Element AI. 2019. 2019 Global AI talent report. Available at https://www.elementai.com/news/2019/2019-global-ai-talent-report. (April 2019).

[7] Danielle Brown. 2018. Google diversity annual report 2018. Available at https://static.googleusercontent.com/media/diversity.google/en//static/pdf/Google_Diversity_annual_report_2018.pdf. (2018).

[8] Andy DiPaolo. 1999. Stanford learning: Worldwide availability on-demand at Stanford Online. *THE Journal (Technological Horizons In Education)* 27, 5 (1999), 16.

[9] Graham Dove, Kim Halskov, Jodi Forlizzi, and John Zimmerman. 2017. UX Design Innovation: Challenges for Working with Machine Learning As a Design Material. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*. ACM, New York, NY, USA, 278–288. https://doi.org/10.1145/3025453.3025739

[10] Brent J Evans, Rachel B Baker, and Thomas S Dee. 2016. Persistence patterns in massive open online courses (MOOCs). *The Journal of Higher Education* 87, 2 (2016), 206–242.

[11] John H Falk and Lynn D Dierking. 2010. The 95 percent solution. *American Scientist* 98, 6 (2010), 486–493.

[12] Geoffrey A. Fowler. 2013. An Early Report Card on Massive Open Online Courses. *Wall Street Journal* (Oct 2013). https://www.wsj.com/articles/an-early-report-card-on-massive-open-online-courses-1381266504?tesla=y

[13] Barbara J Grosz, David Gray Grant, Kate Vredenburgh, Jeff Behrends, Lily Hu, Alison Simmons, and Jim Waldo. 2019. Embedded EthiCS: integrating ethics across CS education. *Commun. ACM* 62, 8 (2019), 54–61.

[14] Philip J Guo, Juho Kim, and Rob Rubin. 2014. How video production affects student engagement: An empirical study of MOOC videos. In *Proceedings of the first ACM conference on Learning@ scale conference*. ACM, 41–50.

[15] Tim N Höffler and Detlev Leutner. 2007. Instructional animation versus static pictures: A meta-analysis. *Learning and instruction* 17, 6 (2007), 722–738.

[16] HubSpot. 2017. What is Artificial Intelligence (or Machine Learning)? https://www.youtube.com/watch?v=mJeNghZXtMo

[17] Richard E Mayer and Joan K Gallini. 1990. When is an illustration worth ten thousand words? *Journal of Educational Psychology* 82, 4 (1990), 715–726.

[18] Ipsos MORI. 2017. Public views of machine learning: Findings from public research and engagement conducted on behalf of the Royal Society. (2017).

[19] Andrew Ng. 2019. Machine Learning. https://www.coursera.org/learn/machine-learning

[20] Chris Nicholson. 2015. A Beginner’s Guide to Neural Networks and Deep Learning. https://skymind.ai/wiki/neural-network

[21] Laura W Perna, Alan Ruby, Robert F Boruch, Nicole Wang, Janie Scull, Seher Ahmad, and Chad Evans. 2014. Moving through MOOCs: Understanding the progression of users in massive open online courses. *Educational Researcher* 43, 9 (2014), 421–432.

[22] S.M. West, M. Whittaker, and K. Crawford. 2019. Discriminating Systems: Gender, Race and Power in AI. Available at https://ainowinstitute.org/discriminatingsystems.pdf. (2019).

[23] Tony Yiu. 2019. Understanding Neural Networks. https://towardsdatascience.com/understanding-neural-networks-19020b758230