MAI4CAREU | Master programmes in Artificial Intelligence 4 Careers in Europe

University of Bologna

Computational Ethics

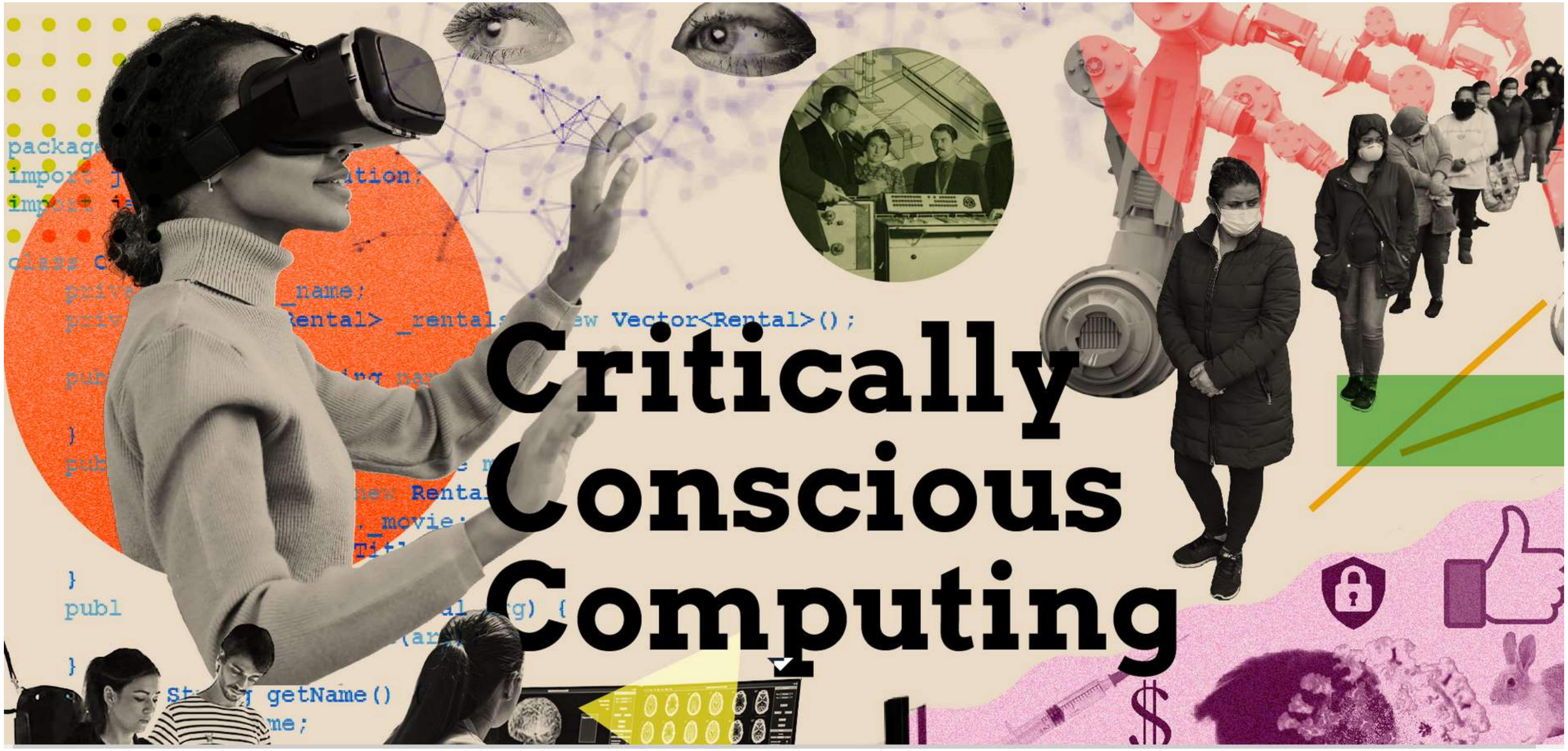
Daniela Tafani

2022/2023 – Second Semester

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action
No 2020-EU-IA-0087, co-financed by the EU CEF Telecom
under GA nr. INEA/CEF/ICT/A2020/2267423

Lecture 4 – Learning material on The power and perils of AI (Extract from *Critically conscious computing*)



Co-financed by the European Union
Connecting Europe Facility

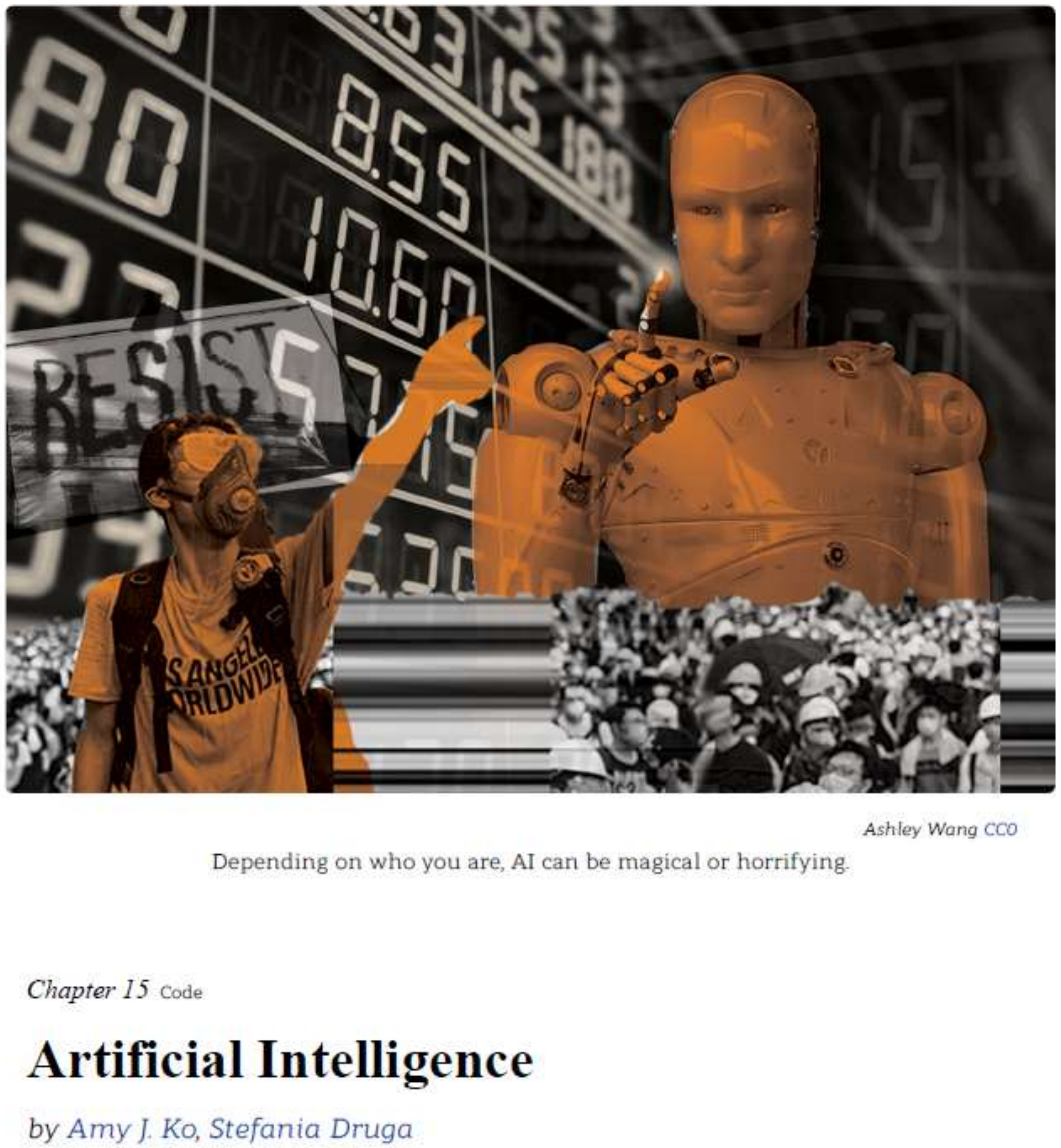
This Master is run under the context of Action
No 2020-EU-IA-0087, co-financed by the EU CEF Telecom
under GA nr. INEA/CEF/ICT/A2020/2267423

Citation

Amy J. Ko, Anne Beitlers, Brett Wortzman, Matt Davidson, Alannah Oleson, Mara Kirdani-Ryan, Stefania Druga, Jayne Everson (2022). *Critically Conscious Computing: Methods for Secondary Education.* https://criticallyconsciouscomputing.org/, *retrieved 9/21/2022.*

License

Creative Commons Attribution-NoDerivatives 4.0



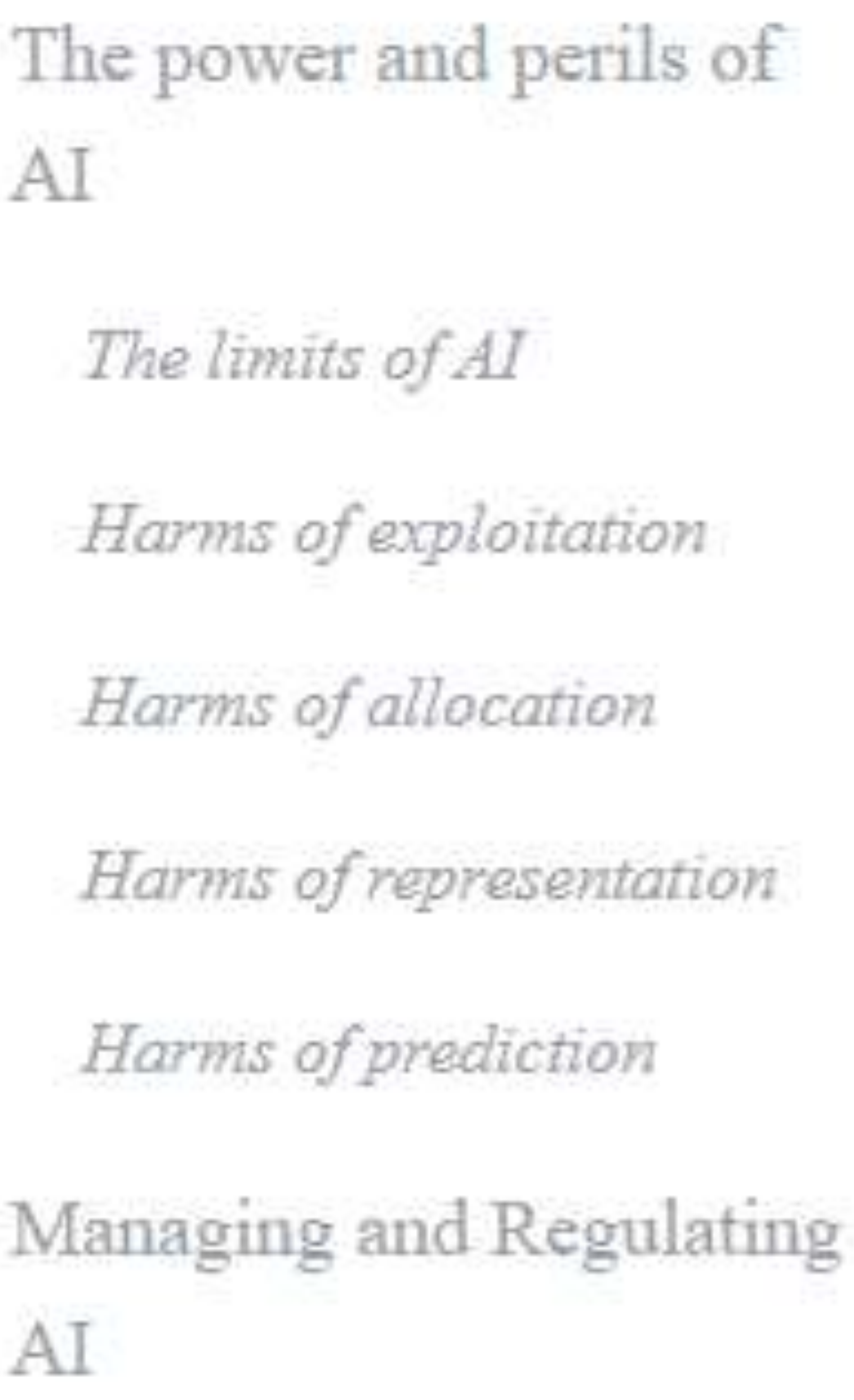
Ashley Wang CC0

Depending on who you are, AI can be magical or horrifying.

Chapter 15 Code

Artificial Intelligence

by *Amy J. Ko, Stefania Druga*



The power and perils of AI

The limits of AI

Harms of exploitation

Harms of allocation

Harms of representation

Harms of prediction

Managing and Regulating AI

Key ideas

- * Artificial intelligence (AI) is concerned with replicating human intelligence and ability, and is increasingly used in software applications, consumer devices, and every major industry to automate and inform decisions.
- * AI has made many advances in replicating specific human abilities, primarily using large data sets created through human labor, and using that data to build machine-learned classifiers.
- * AI perpetuates whatever values and biases were encoded in its algorithms and data, and depending on how they are used, create new systems of oppression.

Ashley Wang CC0

AI can do harm, often unintentionally.

The power and perils of AI

The innovations above can be quite impressive – how can we resist the futuristic wonders of smart assistants, autonomous cars, and acrobatic robots? And with AI increasingly embodied, with voices and physical forms designed to mimic human behavior and leverage human social cues, we increasingly see AI as more than just data and algorithms: we *like* them, and come to see them as human-like, even though they are just data and algorithms[28]. And yet, the problematic applications of machine learning we portrayed in the previous sections make clear that this power is not necessarily good. In this section, we discuss the limits of this power, and some of the many social consequences of misapplying it.

[28] Byron Reeves & Clifford Nass (1996). The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places. *Cambridge University Press*

The limits of AI

One of the most salient findings from AI research since the 1950's is that strong AI is far out of reach[6]. The closest that researchers have come is to create programs that can fool humans into believing they are intelligent for a brief time. Alan Turing described what is now known as the "Turing test" or the "imitation game", in which a human writes text messages to an AI and receives text replies, having a conversation. If the human is fooled for some period of time, one might judge the AI as having achieved some level of mimicry of intelligence. Many have taken the test literally, creating competitions in which people write AIs and compete to see who can trick the largest number of people, or prizes to advance the capabilities of AI in consumer contexts.

But even the best efforts in these competitions and in research reveal the fragile seams in strong AI attempts. All of the techniques above capture *some* essence of human ability or patterns in society, but this is often only true in perfectly calibrated circumstances, and for a very narrow range of human abilities. Applications of AI therefore continue to be *weak* – carefully constructed for particular contexts, and brittle when applied outside those contexts, just like any other kind of computer program.

[6] Adriana Braga & Robert K. Logan (2016). The Emperor of Strong AI Has No Clothes: Limits to Artificial Intelligence. *Information*

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action
No 2020-EU-IA-0087, co-financed by the EU CEF Telecom
under GA nr. INEA/CEF/ICT/A2020/2267423

Harms of exploitation

As weak AI has become ubiquitous in industry applications, primarily through machine learning, data has become a precious commodity: computers can only recognize and produce human speech, label objects in images, and play challenging games like go and chess because companies have amassed large amounts of labeled training data. Many of these data sets are generated entirely through human labor. Sometimes the data is gathered from commercial services without our awareness, such as the way that search engines passively record logs of our queries, labeling the results we select, and using that data set with machine learning algorithms to improve search results. Other data is gathered with human consent, such as when a website login page asks you to select parts of an image that contain traffic lights or stop signs, which are then used to train driverless cars. Some data comes from paid labor, such as the image labels that Facebook gathers from its tens of thousands of low-wage, low-status content moderation staff in developing countries[30]. Many of the AI-fueled conveniences of expensive smart devices are thus built on the backs of no or low wage people, often without consent, and that data is often used against them to surveil and police them[26].

The need for data also links to the issues of privacy and surveillance capitalism we discussed in Encoding Information. It's not just that companies need data in order to train AI to offer new conveniences, it's that they need data to make money, as their primary revenue streams are from selling advertisements[36]. Data, and more fundamentally the invasion of privacy, often without consent, is therefore foundational to the economic systems that drive the creation of AI.

[26] Catherine O'Neil (2016). Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy. *Crown*

[30] Sarah Roberts (2014). Behind the Screen: the Hidden Digital Labor of Commercial Content Moderation. *University of Illinois at Urbana Champaign*

[36] Shoshana Zuboff (2019). The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power. *PublicAffairs*

Harms of allocation

All algorithms, when inserted into decisions about who gets a particular resource, and who does not, can disproportionately harm some groups and not others[3]. AI, however, has been found to be particularly systematic in its discrimination. For example, consider Black Americans, who because they are more likely in poverty, are increasingly denied access to housing, food, loans, and insurance due to applications of AI that are racially biased in their risk predictions[13]. These harms of allocation stem from multiple levels: the algorithms, which prioritize accuracy at the aggregate level, without considering impacts on marginalized groups; data that often underrepresented marginalized experiences; and applications of AI that stem from political goals of reducing fraud rather than helping people in need.

In contrast, other applications of AI often enrich and liberate groups that already have power and wealth. For example, the realities of weak AI such as machine learning that aim to make our lives easier and safer – often for the benefit of the wealthy, and at the expense of marginalized groups. Algorithms that optimize the availability of rideshare services, for example, only benefit those who can afford expensive rideshare trips, and at the expense of drivers who often get paid less than minimum wage after accounting for expenses. AI, then, just like any other code, is often deployed as a tool of wealthy, dominant groups to accrue power, increase wealth, and maintain the matrix of oppression that erases diversity, denies equity, and shuns equality[4].

[3] Ruha Benjamin (2019). Race after technology: Abolitionist tools for the new jim code. *John Wiley & Sons*

[13] Virginia Eubanks (2018). Automating inequality: How high-tech tools profile, police, and punish the poor. *St. Martin's Press*

[4] Abeba Birhane, et al. (2021). The Values Encoded in Machine Learning Research. *arXiv*

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action
No 2020-EU-IA-0087, co-financed by the EU CEF Telecom
under GA nr. INEA/CEF/ICT/A2020/2267423

Harms of representation

While harms of exploitation and allocation are more direct, AI also causes more subtle, deeper harms in the form of misrepresentation, stereotyping individuals in harmful ways. Such harms are often quite simple. For example, machine learning that uses a binary gender feature – male or female – works fine for the cis majority, but erases people with non-binary or gender non-confirming identities. Such data, used by companies like Facebook to determine AI-based ad recommendations, excludes countless people from participating in commerce. Or, consider machine learning that used in face detection algorithms by police: because such machine learning is less effective for Black faces, there is been a surge in false arrests of Black people, spending time in jail, losing jobs and losing money on lawyers to prove their innocence.

Some of these harms stem from a lack of recognition of human, social, and cultural diversity. For example, what is “common” in common sense AI is culturally and socially determined. It’s only obvious to a person familiar with basketball that the spherical object is a basketball; to children who have never seen the game, or people in cultures unfamiliar with the Western sport, such facts would not be obvious. Therefore, the values in common sense AI are highly sensitive to the data on which they are based. The same is broadly true of all AI: because most AI systems are made by English speaking people in Western cultures, and authored by wealthier people with access to computers and the Internet, AI often reflects assumptions about what AI is for, and who it is for.

Some harms, however, stem from AI algorithms themselves, all of which optimize for common cases over edge cases. Anyone part of a group that is less common than the majority is likely to be misrepresented by the logical rules or statistical patterns used to define AI behavior, necessarily resulting in worse accuracy for people in minority groups. Some of these harms stem from data used to train AI, where historical norms or trends – such as dated Western notions about binary gender – end up erasing individual identities. And some of these harms simply stem from how AI is applied in society: for example, it is possible to build less racially biased facial recognition technology, but it would still be used to disproportionately surveil and police Black people in the U.S.

edge case: Any consideration an algorithm needs to make that is not part of the common case, such as unexpected inputs, rare scenarios, or contexts that violate its assumptions about the world.

In some ways, the very notion of categorization, classification, and labeling is harmful. Consider, for example, our applications of machine learning on tardiness earlier. It's tempting to view lateness as a binary construct: students are either late to class or they are not, right? Of course, as any teacher knows, lateness is not an inherently binary phenomenon, because of the many exceptional circumstances that might conspire to make lateness ambiguous. Daylight saving time, late buses, clocks that differ by more than a few minutes, a faulty passing period bell, an emergency like an earthquake, poorly accommodated physical disabilities out of a student's control – all of these circumstances and more erode the idea that lateness is a binary. And the same complexities arise when we try to define many natural and social phenomena, such as race, gender, poverty, wealth, and so on. Supervised learning algorithms that model some aspect of nature or society often must reduce these ideas to a smaller set of discrete or binary values in order to work – but they do so at the cost of accurately modeling diversity. And more so, the people that are erased by the reductive nature of classification are usually people at the margins, who have exceptional circumstances, identities, bodies, and behaviors. Grouping, labeling, classifying, categorizing – all of these are just synonyms of stereotyping, which erases the nuance of difference and diversity.

Harms of prediction

One of the most common applications of machine learning, and statistics in general, is to make predictions about the future. As anyone knows, a prediction is not necessarily going to be true. And yet, many applications of AI frame prediction as if it are trustworthy, fair, and accurate.

Consider, for example, the problem of **recidivism**, which is the tendency for people who are convicted of crimes to commit further crimes[35]; research on recidivism suggests that repeat offenses are primarily caused by the fragile social supports and stigmas that previously incarcerated people face after release. When judges make sentencing decisions after a jury convicts them of a crime, they often make recidivism predictions, assessing how likely the convicted person is to commit another crime. If they decide the likelihood is high, they may give the person a longer sentence, hoping that separating a person from society will protect society from their crimes; if it is low, they may give a shorter sentence. There's obviously immense opportunity for bias in these human judgment, as the judge's prediction of recidivism might be swayed by the color of the convicted person's skin. And so some software companies saw an opportunity to create software for courthouses that would use machine learning to make these predictions instead, hoping to remove the potential for a judge's bias by replacing their judgment with supposedly "neutral" AI.

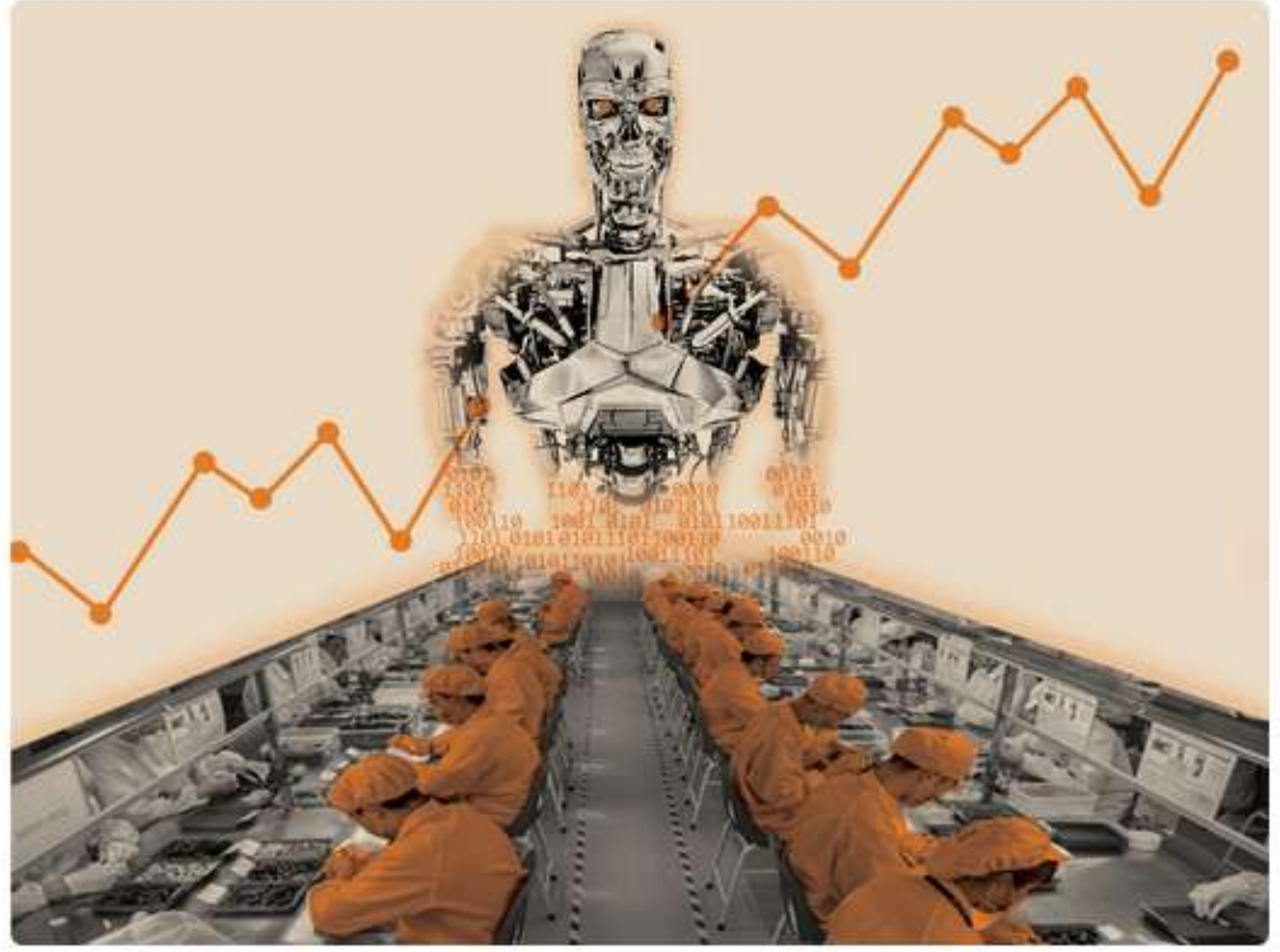
The most popular software, COMPAS (Correctional Offender Management Profiling for Alternative Sanctions), launched in 1998, used the approach of taking a large data set of past convicted people, gathering 137 different factors about the accused (the degree of the crime, whether it was a drug offense, the offenders' age), but did not include race as a factor. Machine learning algorithms, as we will describe below, build statistical models that find how much to weigh different factors to make the most accurate prediction possible; they use data on past convictions to "train" a classifier, then use that trained classifier to make predictions about new cases. COMPAS did this same thing, using a historical record of convictions and re-convictions to make predictions about people newly convinced. Despite COMPAS's insistence that race was not included as factor, the factors that it did include were nevertheless racially biased: Black men are more likely to be arrested for crimes because they are surveilled and policed more than other identities; they are more likely to be convicted of crimes, even when they are later found to be not guilty. The machine learned classifier in COMPAS learned these racially-biased patterns, and reproduced them, without ever explicitly including race as a factor. In fact, research showed

[35] Edward Zamble & Vernon L. Quinsey (2001). The Criminal Recidivism Process. *Cambridge University Press*

that the COMPAS machine learned classifier was just as biased as lay people with little or no criminal justice expertise, and in fact less accurate and more biased than experienced judges[9]. And yet, because of a misplaced faith in the neutrality and objectivity of computers, judges across the United States have deferred to COMPAS and its racially biased predictions to make their sentencing decisions, rather than using their knowledge of the individual accused. The result is that judges using COMPAS give Black men even longer sentences than other racial groups.

The developers of COMPAS did not evaluate these outcomes before they built it; their priority was selling software to courthouses, and cities' priorities were reducing perceptions of bias (instead of actual bias, or more importantly, the larger systemic bias in policing, prisons, and poverty). No one actually investigated this bias until the public started demanding audits of these algorithms and researchers conducted these audits, which revealed their ineffectiveness and bias. This application of weak AI, therefore, prioritized certain people's concerns (judges, politicians, and the White majorities that elected them) at the expense of others (Black men).

[9] Julia Dressel & Hany Farid (2018). The accuracy, fairness, and limits of predicting recidivism. *Science Advances*



Ashley Wang CC0

AI has bias, but it's not always clear what to do about it.

Managing and Regulating AI

Because AI, and computing in general, has such great potential for harm, there are numerous approaches to trying to mitigate its risks.

Transparency. One approach is to advocate for transparency – the degree to which people other than an AI's creators can see the data on which AI is based and the algorithms it uses to process it[1]. Without transparency, the public is limited to observing the patterns in AI behavior, without being able to see their underlying causes. Thus far, the world has been so enamored with the possibilities of AI that it has asked for little transparency – most private companies view their data and AI algorithms as valuable trade secrets – and only some governments have committed to transparent uses of AI. Just as when code is kept secret, when the data and AI algorithms used are kept secret, there is no way to build public trust in their underlying logic.

Explainability. Even when AI is transparent, the behavior of AI algorithms can be much harder to explain than a computer program[27]. After all, a program has logic, written in a programming language, that someone familiar with that language can read, analyze, and make conclusions about with high confidence. Comprehending a program's behavior is by no means easy, but it is feasible with time and expertise. AI, and especially machine learning, in contrast, often have little explicit logic. Their behavior is inherently determined by the data selected, the statistical analyses applied to that data, and the emerging patterns in that data. The answer to "*Why was my loan application rejected?*" for a computer program might be "*The program inspected your responses to these three questions and based on our rules, you are not eligible*", but the answer for a machine learning based prediction is "*Because of the historical records of hundreds of millions of Americans, and their patterns of repayment, we predict that you will be likely not to repay.*" In fact, ask any machine learning programmer to explain the output of a classifier, and the best answer they can give is often "*Because of all the data.*"

Regulation. Because AI behavior can be so hard to explain, even when it is transparent and the harm is measurable, a central question in AI is "who is responsible for AI behavior?" Is it the people who use the AI that someone created? Is it the people who created the AI? The people who chose the data on which the AI was based? Is it the people who gathered the data that shaped its behavior? These questions are perhaps most salient in the case of driverless cars, which use AI extensively: when autonomous vehicles kill someone, who is to blame? There are few broadly accepted regulatory frameworks that answer these questions. Many Black scholars have advocated for regulation on applications of facial recognition technology; other scholars are calling for even more pervasive debates about law over the coming decades about AI culpability[33]. Increasingly, whistle blowers from industry are even calling for companies who apply AI to be held accountable to the harms they cause.

[1] Mike Ananny & Kate Crawford (2016). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society*

[27] Emilee Rader, et al. (2018). Explanations as Mechanisms for Supporting Algorithmic Transparency. *ACM SIGCHI Conference on Human Factors in Computing Systems (CHI)*

[33] Jacob Turner (2018). Robot Rules: Regulating Artificial Intelligence. *Springer*

Ethics. When asking questions about agency, intelligence, causality, and blame, questions about AI quickly turn to questions of ethics, morality, and justice. Why are we creating AI? Who does it benefit? At what cost? Do we want artificially intelligent things in society? What kind of society do we want? And what is the cost of putting so much time and attention into managing the power of AI when so many people in the world are still struggling with such basic needs, such as food, clean water, shelter, and safety? Currently, these are questions largely being answered by large, wealthy, monopolistic, American companies like Google, Amazon, Apple, Facebook, and Microsoft. If we want a democratic society that reflects all of our views of a just society, then everyone needs the literacy above to ponder these questions and shape policy preferences[17].

[17] Anna Jobin, et al. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*