AI Human Impact
Toward a Model for Ethical Investing in AI-Intensive Companies

Working Draft 10 July 2020
James Brusseau
Philosophy
Pace University, New York City
jbrusseau@pace.edu

Abstract

Does AI conform to humans, or will we conform to AI? An ethical evaluation of AI-intensive companies will allow investors to knowledgeably participate in the decision. The evaluation is built from nine performance indicators that can be analyzed and scored to reflect a technology's human-centering. When summed, the scores convert into objective investment guidance. The strategy of incorporating ethics into financial decisions will be recognizable to participants in environmental, social, and governance investing, however, this paper argues that conventional ESG frameworks are inadequate for AI-intensive companies. To fully account for contemporary technology, the following categories of evaluation will be developed and featured as vital investing criteria: autonomy, dignity, privacy, performance. With these priorities established, the larger goal is a model for humanitarian investing in AI-intensive companies that is intellectually robust, manageable for analysts, useful for portfolio managers, and credible for investors.

Keywords

AI, Finance, Ethical investing, AI ethics, ESG

Traditional ESG Does Not Work for AI

Traditional Environmental, Social, and Governance (ESG) rating was forged for the industrial economy with its hazardous working conditions and polluting machines (MSCI 2020). Artificial Intelligence companies do not fit in.

The disconnect is material – cement and smokestacks diverge from pixels and digital exhaust – but the significant divide is human. Standard ESG categories and actions gravitate around *collectives*. Fair trade organizations rally for farmers in developing nations (Mason 2016), unions organize for women's rights on the work floor (Eccles 2018: 16), environmental advocates promote cleaner water for future generations. These social and political movements *unify* activists, which means detailed personal information about specific participants is unnecessary. It is even a distraction because the project is to suppress individual differences in the name a common cause.

Artificial intelligence moves in the other direction: everything starts with personally identifying information. The proposal to *Nudge for Good* (Borenstein and Arkin 2017) models this new reality. Users' memories, vulnerabilities, and urges are gathered within a big data pool and analyzed with predictive algorithms to create micro-targeted solicitations for charitable causes. These messages are crafted for the psychological profile of one identified person, not for group appeal. They are delivered to a specific Facebook user, or voiced by a single household robot, not announced on indiscriminate public media. (Borenstein and Arkin 2017: 501-502). Shoshana Zuboff has described an emergent ecosystem of data, algorithms, and details of public and private lives. They combine as behavioral futures markets, places where knowledge can be purchased about one person, and where they will be, at what time, in what mood (Zuboff 2019: 8). As for how that information will be used, the question remains open. What is certain, however, is that across the technological and economic spectrum, an inversion is occurring: the human condition is no longer defined by the unifying elements of collectives, but by the individualizing particularities of users.

The inversion explains why privacy concerns have become so pressing in public conversations and corporate meeting rooms (West 2019). It also means that the most tangible socio-economic threats no longer come from outside of ourselves, they are no longer rigid social customs or imposing governmental regulations. Instead, the immediate peril is *our own* dataset, it is the information defining who we are – our habits, tastes, fears, desires and aspirations – that may be engineered to provide gratifying experiences and opportunities, but that can also be twisted to control where we go and what we do.

The paradigmatic theoretical case is predictive policing because of the question it asks: *Is my data innocent or guilty, liberating, or confining?* Will the personal information that has been gathered about me invigorate my life, or restrict it? Whether the AI is stationed in a police station, or on the LinkedIn career platform, or the OKCupid romance site, or at an airport security kiosk, or inside a hospital emergency room, the question is the same.

Because the question about whether AI serves humanity, or humanity serves AI fundamentally asks whether the data and algorithms vitalize or debilitate on the level of single individuals, the first metric for responsible investing will be autonomy: does a technology expand self-determination? The individualizing values of dignity and privacy follow as key performance indicators. Conventional responsible investing metrics will also be included in the evaluation, ones recognizable to ESG investment strategists. But what makes AI humanitarianism different – and what requires a new and distinct model for ethical investors in AI-intensive companies – is evaluation that begins with persons, not people.

Overview of AI Human Impact

AI is increasingly deployed to facilitate conventional ESG investing (Antoncic 2020), but little has been done for the reverse: establishing standards for nonfinancial performance of companies employing AI at the core of their operation. We have AI for ESG, but not ESG for AI.

Still, much of what will be required to produce a human-centered investing blueprint has already been accomplished. Since 2017, more than 70 AI and big data ethical principles and values guidelines have been published (Jobin et al. 2019:3; Fjeld et al. 2020; Hagendorff 2020), and even as the wave crested, researchers were already mapping the overlaps, forming principles of the principles (Floridi and Cowls 2019, Hagendorff 2020). Because many are ethical in origin, the contributions tend to break along the lines of academic applied ethics (Mittelstadt 2019). Corresponding with libertarianism there are values of personal freedom. Extending from a utilitarian approach there are values of social wellbeing. And, on the subject of responsibility, there are values focusing on trust and accountability for what an AI does.

Each of the mainstream collections of AI ethics has their own way of fitting onto that trilogical foundation, but the *Ethics Guidelines for Trustworthy AI* sponsored by the European Commission (AIHLEG 2019) is representative (Clement-Jones 2019), as is the *Opinion of the Data Ethics Commission* of Germany (DEC 2019). They are arranged in the figure below for comparison, along with the values grounding AI Human Impact: autonomy, dignity, privacy in the personal freedom group; fairness, solidarity, sustainability in the social wellbeing group; performance, safety, accountability in the technical trustworthiness group.

Figure 1

	AI Human Impact	***Ethics Guidelines for Trustworthy AI*, European Commission**	***Opinion*, German Data Ethics Commission**
Personal Freedom	**Autonomy**	**Human agency and oversight**	**Self-determination**
	Dignity		**Human dignity**
	Privacy	**Privacy and data governance**	**Privacy**
Social Wellbeing	**Fairness**	**Diversity, non-discrimination, fairness**	
	Solidarity		**Justice and Solidarity**
	Sustainability	**Societal and environmental wellbeing**	**Sustainability**
			Democracy
Technical Trustworthiness	**Performance**		
	Safety	**Technical robustness and safety**	**Security**
	Accountability	**Accountability**	
		Transparency	

There are discrepancies, and some are superficial. The E.C. *Guidelines* split 'Accountability' and 'Transparency,' whereas AI Human Impact unites them into a single category. The German Data Ethics Commission joins 'Justice' and 'Solidarity,' whereas AI Human Impact splits them into 'Fairness' and 'Solidarity.' Another difference is more revealing. Performance as an ethical principle only occurs in the AI Human Impact model because it is extremely important to investors: the *reason* to get financially involved in the first place is to make money. Not only to make money, but that is the initial step.

Consequently, how well the machine performs is a critical concern: an AI that cannot win market share will not have human, or any impact. That consideration may be less pressing for public sector investigators, especially those whose work depends more on institutional grants than economic profit. Also, to the extent that institutions are establishment in nature – to the extent they resist change – their patronage may incentivize defenses of humanity *against* AI innovation, more than humanist contributions *to* it (AIHLEG 2019: 2, DEC 2019: 5). Regardless, because AI Human Impact begins with financial interests, enhancing performance is a constitutive element of the model.

Instead of slowing progress or constraining engineers, ethical AI investing *provides humanistic feedback to catalyze more and faster development*. One part of that contribution is the illumination of risk that could precipitate technology's broad social rejection, as around facial recognition technology currently (Press release 2019). The more significant contribution, however, is to orient AI design toward individual human potential. The project is to describe how data and machine learning can be *measurably* converted into vital experiences that replace the numbing and banal activities now consuming too many human hours. Accelerating and multiplying personal opportunities, that is the purpose of AI human impact investing. The financial premise is that humanitarian purposiveness yields outstanding returns in the medium and long term.

The following sections develop nine categories for analysis and scoring. Individually, they reveal discrete aspects of a company's ethical performance. Then, a formula for summing the scores will be proposed to calculate a broad human impact rating for investments in AI-intensive companies.

Personal Freedom Metrics: Autonomy, Dignity, Privacy

Autonomy means giving rules to oneself, which stands conceptually between living by imposed rules, and the senseless chaos of life without any rules at all. In lived AI experience, autonomy exists when data and algorithms help us act for our own reasons.

Because no company currently reports on their autonomy bottom line, this performance indicator proves difficult to abstract from corporate disclosures and public information sources. Still, a resourceful analysis can distinguish those AI companies that constrict users' self-determination, from those that expand it.

Constrictors may employ *dark patterns*, interfaces surreptitiously prompting decisions that users might not otherwise make (Narayanan et al 2020). As Facebook's founding President explained:

> How do we consume as much of your time and conscious attention as possible? We need to give you a little dopamine because someone Liked or commented on a photo or a post, and that's going to get you to contribute more content, and that's going to get you more Likes and comments. It's a feedback loop exploiting a vulnerability in human psychology (Allen 2017).

The psychological vulnerability is chemical, and activated by social media *Likes* dosed by algorithms. The result is insidious: users are confined by *their own data*, and in two senses. First, the material that keeps drawing them back is their own posting. Second, the dopamine is calibrated to the users' personal profile: if the Likes are too many for that specific person, or too few, interest dissipates (Remia 2015). In the end, it may even be that users *enjoy* being constrained by their own posts and personal information, but the autonomy score is negative.

Another dark pattern is AI nudging, defined as controlling behavior without relying on legal or regulatory mechanisms (Thaler and Sunstein 2008). This behavioral modification has been modelled for use in home chatbots to "foster empathy that nudges a user towards performing charitable acts." (Borenstein and Arkin 2017: 502) The key is information gathering for surgically precise appeals. "If the user has a family history of a particular illness like heart disease, the robot could suggest contributing to an associated charity, like the American Heart Association." (Borenstein and Arkin 2017: 503) So, a sad episode converts into a donation, and while no one is against charity, the gift fails to cancel what happens on the level of autonomy: users are not receiving

data to help them make decisions so much as receiving decisions generated by their data. The score is negative.

Autonomy can also be constricted through dependence (AIHLEG 2019: 16). In medical domains including ophthalmology, oncology and dermatology, machine learning algorithms have outperformed human counterparts in competitive tests of detecting diseases from clinical images and, in these contexts, it may become difficult for doctors to trust their own learning more than the machine's (Grote 2019). Subsequently, deference to the statistically proven performance may become habitual: if the AI is going to be right, why bother to think through a diagnosis independently? The result is deskilling, dependence, and an adverse affect on self-determination.

How does an autonomy score turn positive? With AI designed to facilitate human *doing*. While it is true that Facebook can trap users in their own banality, the platform also allows entrepreneurs to establish a business, advertise online, and be serving clients in hours. And, while AI image analysis may render doctors redundant, it can also heighten performance by drawing attention to anomalies that may have otherwise been missed. With respect to any specific AI-intensive company, the root autonomy question – *Does the AI open opportunities or close them?* – may not yield a binary answer, and so require careful weighing of how the AI works on balance in the flesh and blood world.

One unambiguous way that AI supports autonomy is by catalyzing experimentation. Part of the essence of human freedom is the ability to try new things, and AI contributes when helping users access possibilities – professional, romantic, cultural, intellectual – outside the funnel of those already established by their habits. The challenge is to mechanically produce serendipity.

As anyone who has scrolled Netflix hoping for a movie suggestion that is unfamiliar to established tastes, but also enjoyable, has learned, serendipity is hard. Part of the problem is the way Netflix engineers use AI to predict satisfying recommendations. In a public talk, a company engineer spent nearly

his entire period describing techniques to isolate new films that significantly *matched* those specific viewers had already liked. The strategies include extrapolating from what the client has already seen, to finding other viewers who resemble the target subject, and checking what *they* like. Not until the final sentences of the presentation did he reveal an effort within Netflix to help viewers escape the logic of similarity and discover new, unexpected possibilities:

> Production biases in ML models can cause feedback loops to be reinforced by the recommendation system. We do research and development in the causal recommendation space specifically to get our recommender models out of this feedback loop (Deoras, Anoop 2020).

It was a tantalizing conclusion, but also frustrating as no details were provided.

For autonomy scoring, those last sentences merit follow-up. *Can Netflix provide users with film suggestions that their users could not have foreseen wanting?* If so, the machine is bettering human recommenders. It is also creating new opportunities, expanding self-determination.

AI serendipity means helping users escape the trap of their own accumulated data. Statistical work addressing the challenge is currently underway in the area of social media polarization (Celis 2019: 160). Online users reliably maintain interactions with others who share their beliefs and values, and that can lead to an echo-chamber effect: people's established views feedback and intensify in a confining circle (Bozdag and van den Hoven 2015). One serendipity response constrains selection algorithms to contain examples from imperfectly related ("non-optimal") groups (Celis 2019: 168). Possibly, this intentional error is a step toward mechanically provoking serendipity. More work will need to be done both conceptually and technically, but whether it is Netflix viewers seeking unexpected but delighting movies, or social media participants seeking unfamiliar but provocative connections, the autonomy tension is the same. Big data and predictive analytics can reinforce habitual

experiences, or, they can diverge from – even escape – the narcotic tranquility of the pleasantly familiar. Divergence and escape serve autonomy.

There is also the money question. On one side, tightening personalization in AI provisioning of user opportunities is pervasive online because the efficiency presumably creates higher revenue for the platform (Sakulkar and Krishnamachari 2016). Still, boredom is a human reality, and some research indicates that providing fruitful discovery opportunities for users holds their attention better than simple repetition of what has already proved satisfying (Kamehkhosh et al. 2020). Serendipity, in other words, may dovetail with conventional business incentives.

Summarizing, autonomy as a key performance indicator for AI ethical investing measures whether the AI oppresses or vitalizes self-determination.

Figure 2

AI Ethics of Personal Freedom	
Autonomy: ***Does the AI debilitate or invigorate self-determination?***	**Negative** • **Platforms diminish opportunities** • **AI decides *for* users** • **Dependence fostered** **Positive** • **Platforms open opportunities** • **AI enables user decisions** • **Opportunities for experimentation created**

Dignity, the second personal freedom criterion, requires that people be treated as ends, and not only as a means (Kant 1996: 429). The dignified are subjects and not objects.

Human dignity begins as freedom from exploitation. It precludes being understood as pure data, as material for processing, profit-taking, and deleting because the dignified have their own independent projects intrinsically meriting recognition. Dignity is also freedom from patronization, meaning that taking responsibility for one's own acts is not a burden but a positive

right. The terminal example is the execution of murderers which dignity requires not as an obligation to the original victim, or to society, but as an expression of *respect for the murderer* (Kant 1996: 333). Conversely, failure to execute is not merciful or benevolent but insulting: the criminal is belittled as incapable of personal responsibility.

Like autonomy, the AI contribution to human dignity is difficult to abstract from corporate disclosures and public sources. Nevertheless, careful analysis can distinguish AI companies that treat users as means, from those treating users as ends in themselves.

In 2014, a co-founder of the OkCupid dating site published a blog entry titled *We Experiment on Human Beings!* It has since been taken down, but while visible, at least some of the experiments were recounted (Berinato 2014). In one, users who the OkCupid algorithms determined to be incompatible were told the opposite. When they connected, their interaction was charted by the platform's standard metrics: how many times did they message each other, with how many words, over how long a period, and so on. Then their relationship success was compared against pairs who were judged truly compatible. The test presumably measured the power of positive suggestion: Do incompatible users who are *told* they are compatible relate with the same success as true compatibles? (Rudder 2014)

The answer is not as interesting as the users' responses. One asked, "What if the manipulation is used for what you believe does good for the person?" (Rudder 2014) The appeal here is to the fine print of the dignity requirement: treat others as ends and not *only* as means. In the real world, it can be true that exploiting others mixes with helping them. The question dignity asks is: Which one serves the other? Are the romance-seekers being manipulated in experiments for *their* ultimate benefit because the learnings will result in a better platform and higher likelihood of romance? Or, is the manipulation wholly about the platform's owners and their marginally perverse curiosities?

Part of the answer lies in the fact that the experiment – and therefore the manipulation – was revealed to the users, leaving them free to respond. It is

not clear how many responded by cancelling their accounts (Berinato 2014). In any case, the dignity question for AI is not whether the technology helps users, it is whether users *determine for themselves* what the word "help" means.

The other side of dignity is freedom from condescension. This can be a stern demand. AI chatbots, for example, are increasingly employed to ward off depression, especially among the elderly (Pereira 2019). The chatbots are also increasingly difficult to detect as mechanical (Shestak et al. 2020). Further, patients respond better to interlocuters who they believe to be human (Chan et al. 20187). Those premises write their own conclusion: deceptive AI chatbots should be deployed to elderly patients. The result would likely be diminished depression, but as long as the patients are not informed of the AI impersonation, it remains true that the entire process depends on patronization: the dignity objection is that users are being treated as unworthy of fending for themselves when it comes to their own treatment. So, the problem with deceitful AI is not exactly that decisions are being made *for* patients (that is the autonomy objection), it is the implication that patients cannot manage the deciding. When that happens, the dignity score must be adverse.

As an ethical performance indicator, dignity measures whether an AI respects users' independent projects, and respects users taking responsibility alone for where those projects lead.

Figure 3

AI Ethics of Personal Freedom	
Dignity: *Does the AI respect users' independent projects, and respect users taking responsibility alone for where those projects lead?*	**Negative** • **Users as objects, tools, means to others' ends** • **Users not treated as responsible for actions** **Positive** • **Users as subjects with projects, ends-in-selves** • **Users alone responsible for their data and algorithm decisions**

Privacy is the third personal freedom metric, and defined as control over access to our own personal information (Westin 1968: 3). Because privacy is an ability, not a state, it cannot be measured by how many people know how much about someone. Instead, privacy gauges the power *to determine* who knows how much. Kim Kardashian, consequently, is one of the most private people in the world, which does not mean her personal life is closely kept, but it *is* closely guarded: she strictly controls her own exposure. The fact that she chooses overexposure by conventional standards subtracts nothing from her privacy.

Several years ago, a news report circulated about a woman who lived nocturnally as a sex worker, while maintaining an ordinary daytime identity with an academic email address and typical social media postings. The two worlds kept their distance, until she and her clients began appearing in each other's "People You May Know" recommendations on Facebook (Hill 2017). She tried to turn off the connections with the expected results, and so learned first-hand the difference between degree of intimate availability which has nothing to do with privacy, and *control* over that availability which is privacy.

The reason for privacy – and the reason it exists as a category of personal freedom – is to decide for ourselves who we want to be. Normally, the decision does not swing as far apart as professor and prostitute, but in smaller ways all of us depend on limiting personal information to form definitions of

ourselves as we go through a typical day. Parents display a goofiness in front of their young children that they would be appalled to reveal to their coworkers. The persona many adopt in the workplace would aggravate a spouse, and spouses define each other in unique ways when no one is watching. So, control over access to personal information is not an occasional concern, it is an everyday part of creating an identity: at any given time and in the company of selected others, we determine our own identity by exposing parts of ourselves, and by concealing others. By exercising privacy, we become who we are.

On the practical level of scoring for responsible investing, significant advances have been made. The General Data Protection Regulation (GDPR) is a milestone advance, and departing from Article 5 – *Principles Relating to Processing of Personal Data* – researchers in Swedish and Danish universities have assembled specific criteria in the areas of data governance and cyber security that can measure privacy performance. With slight modifications, they are (Vinuesa et al. 2020b):

- Users control their own data's collection and use.
- Data collected and used transparently.
- Data-minimization principle in effect: only information necessary to perform the AI function is gathered, data storage is local, and temporary.
- Privacy-by-design engineering.
- Security ensured by user authentication to prevent risks such as access, modification, or disclosure of data.
- A cybersecurity yield is available, one that measures the magnitude and efficiency of a company's security expenditure in relation to the value at risk (Nolan et al. 2019).

Within today's ESG research community, the privacy category is well-established. Sustainalytics's *Managing data privacy risk: comparing the FAANG+ stocks* assesses how seven major technology corporations perform in data privacy. Facebook and Amazon are graded as vulnerable to high risk exposure attributable to weak data management. Apple is reported to be well-

positioned due to strong data governing policies (Sustainalytics 2018). Another sector leader, MSCI, benchmarks more than 600 companies annually on risk linked to privacy, with one of their 37 ESG Key Issues titled: Privacy and Data Security (MSCI 2019). The information, that means, required to score AI-intensive companies on their privacy performance is increasingly available.

Summarizing, privacy as an ethical investment performance indicator measures whether an AI adds to, or subtracts from users' control over access to their own personal information.

Figure 4

AI Ethics of Personal Freedom	
Privacy: ***Does an AI add to, or subtract from users' control over access to their own personal information?***	**Negative** • **Personal data collected indiscriminately** • **Opacity about where personal information goes, how used** • **Cybersecurity inadequacy** **Positive** • **User controls personal exposure** • **Transparent data collection and use** • **Data-minimization principle for collection, use, storage** • **High cybersecurity**

Social Wellbeing Metrics: Fairness, Solidarity, Sustainability

Fairness can be understood as applied to individuals, or to identity groups. Applied to individuals, it is traditionally defined: equals treated equally, and unequals proportionately unequally (Aristotle 1934: Book 5:3:13). If two people have similar financial backgrounds and apply for comparable loans, then an AI designed to produce lending decisions should arrive at similar disbursement results. Conversely, to the extent two applicants present unequal

financial strength (different income levels, outstanding debts, and so on), their loan terms should be proportionately dissimilar. For AI designers within this lending model, fairness is straightforward. Data is processed to predict who will – and who will not – repay a loan, and the resulting probability corresponds with a loan application decision. Being fair, consequently, means being accurate.

Fairness can also be understood in terms of identity groups. Frequently advocated under the title of social justice, the idea is not that two people who are similar in terms of their finances are also determined to receive similar credit opportunities, instead, balances are sought between men as a group and women grouped, or between races, or other communities. To explore these fairness concepts, several statistical models have been developed. Equalities in algorithmic lending can be sought, for example, as the fraction of non-defaulting members from racial groups (Hardt et al 2016: 17-19; Narayanan 2018: 00:34:06). In other words, if you take all those individuals who were awarded loans, and divide them into race categories, and then check the proportion of members in each category that repaid, the percentages should be about equal. If they are not, if one racial aggregate contains relatively few defaulting members, that suggests mediocre credit risk applicants in that group are getting rejected, while in other groups mediocre risks are getting accepted as reflected by their relatively high default proportion. So, the ideal of equal opportunity between races may justify reweighing the lending algorithm to bring nonpayment proportions into alignment.

The overall result is a distinction between two fairness views: one is about individuals and treating them equally, while the other concerns groups and treating them analogously. In technical terms, the debate is between calibration (individual accuracy) and parity (group balance), and the fairness dissonance between these two possibilities has been among the highest profile discussions in recent AI ethics, especially involving the Northpointe information services and technology company, and it's model for predicting recidivism risk (Washington 2018; Angwin et al. 2016) .

For the purposes of AI ethical investing, there is no right or wrong here. What counts is awareness that the sides exist: when AI designers opt for calibration, or for parity, they are making a fairness decision. An AI intensive company with a strong fairness score is one that knows where it stands, and why.

Fairness is not only about distribution (of loans, and other opportunities), it is also about *representation*. Google searching "CEO" images turns up overwhelmingly white male faces, which corresponds with the gender and race reality (Lam 2018). The fairness question emerges when that truth crosses this one: people may be less likely to aspire to be a CEO if they do not perceive others like themselves have already followed that route (Barocas and Selbst 2016). Here again arises the dichotomy between fairness resting on accuracy (CEO image search should show truly representative CEO images), and fairness resting on opportunity (CEO image search should be tweaked to invite all of society into the aspiration). This is a true dilemma – Design for accuracy, or for opportunity – but for the scoring of AI ethics, no resolution is required, only awareness that algorithmic weightings are fairness decisions *and* decisions about what counts as fairness.

Another AI intersection with fairness involves the data used for modeling and training applications. It may reflect individual, cultural, and historical biases (Gianfrancesco et al. 2018, Char et al. 2018), and may lead to unwarranted disadvantages in treatment for individuals or groups (Bobrowski and Joshi 2019, Goodman et al. 2018). Examples are numerous. A recent study found that an AI teaching itself English from human writings ended up acquiring human-like prejudices in expression, notably against black Americans and women (Caliskan 2017). An Asian man's passport photo was repeatedly rejected by an AI scan because it read the subject's eyes to be closed (Griffiths 2016). Speech recognition products have been found to be more accurate for male speakers than females (Tatman 2016). None of these examples is obviously indicative of discrimination as intentionally harmful, they instead seem attuned to biases embedded in initial training data. Part of AI fairness, consequently, involves data inspection and cleaning: where did the original information come from? How might non-material factors pollute it? Because

fair outcomes require fair data collection (IEEE 2019:190), an AI ethics score reflects a company's attentiveness to the data gathered before training and deployment.

After deployment, there is the fairness question about bias amplification: unbalanced outcomes can feed back into the process and exacerbate discrimination. At Amazon, a workforce where men were superior in terms of quantity infected training data employed to algorithmically rate job applications with the lesson that male employees were superior in terms of quality. Female resumes were correspondingly downgraded (specifically those featuring graduation from two women's colleges, as well as various other word combinations, including "women's chess club") (Kodiyan 2019: 2). The AI was rewritten, and then discontinued, but the risk persists that algorithms can start with uneven information, recycle it, and repeat as the outcome tilts ever further out of balance.

Finally, the IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems adds that a diverse workplace may help AI designers encounter and remedy bias problems, both in the initial data and in subsequent iterative processing. Diversity in this context refers not only to traditional identity groups, but also educational and professional backgrounds: interdisciplinary teams might include computer scientists along with experts in medicine, architecture, law, philosophy, psychology, and cognitive science (IEEE 2019: 126, 19). To the extent that is correct, workplace diversity itself can be gauged as a proxy for protection against the risk of unfairness.

Condensing the discussion, fairness is scored as equal treatment for society's members. Key elements of the performance indicator include:

- AI designers are knowledgeable with respect to the accuracy versus opportunity debate at the core of AI fairness.
- Safeguards erected against biased (in social and statistical senses) data.
- Mindfulness of bias amplification in AI applications.
- Workplace diversity as a proxy for protection against the risk of unfairness.

Figure 5

AI Ethics of Social Wellbeing	
Fairness: ***Are all treated equally?***	**Negative** • **No engagement in accuracy versus opportunity debate** • **Failure to account for bias embedded in data, or exacerbated in processes** • **Minimal workplace diversity** **Positive** • **Engagement in accuracy versus opportunity debate** • **Safeguards against bias embedded in data and exacerbated in processes** • **Genuine workplace diversity**

Solidarity, the second criterion of social wellbeing, is the inclusiveness of no one left behind (Microsoft 2018). In AI medicine, because the biology of genders and races differ, there arises the risk that a diagnostic or treatment may function well for some groups while failing for others (Noor 2020; Wang Keng 2018). The ethical difficulty is captured by a hypothetical: If an AI scan analysis for Melanoma is trained on data from white men, and it proves effective, should white males who may have the disease wait to use the technology until data has been gathered and training administered for all races? A strict solidarity posture could respond affirmatively, while a flexible solidarity would allow use to begin so long as data gathering for unrepresented groups also initiated. Solidarity's absence would be indicated by neglect of potential users, possibly because a cost/benefit analysis returns a negative result, meaning some people get left behind because it is not worth the expense of training the machine for their demographic segment.

Another solidarity element is a Max/Min distribution, one where the benefits of an AI distribute maximally to those who have least (Rawls 1971: 266). One way to measure AI benefits is as wealth, so the question about benefits distribution converts into this: Does an AI tilt advantages toward those with the least resources? In some cases, a positive response seems likely.

Psychological carebots designed to help fight depression may be more affordable and accessible than human, face-to-face treatment. Their development, that means, provides a mental health advantage to some who previously could not afford it (Singh 2019). By contrast, there are indications that AI as an industry is exacerbating inequalities instead of remedying them. According to one thorough study, "Artificial Intelligence puts more low-skilled jobs at risk than previous waves of technological progress" (Nedelkoska and Glenda 2018). It may result, of course, that job losses in one sector create better or new opportunities elsewhere. Regardless, on the microeconomic scale and with respect to rating specific AI products and companies in terms of Max/Min, a positive score goes to those bringing the greatest benefits to users who have the least.

Figure 6

AI Ethics of Social Wellbeing	
Solidarity: ***Is everyone included, with the most going to those who have least?***	**Negative:** • **Limited access to AI** • **Benefits clustered among the most advantaged** **Positive** • **Inclusiveness prioritized** • **Most benefits distributed to the least advantaged**

Sustainability is the third criterion of social wellbeing, and it applies a time horizon to nonfinancial performance: the question addressed to an AI-intensive company asks whether its products serve society's flourishing over the medium and long term. To establish metrics, the United Nations 17 Sustainable Development Goals (SDGs) serve as convenient scoring silos (United Nations 2015) because they are well known to ESG researchers and already feature prominently in their investing strategy (MSCI 2019b). For even greater granularity, the seventeen have been broken into more than 200 sub-indicators by the UN Inter-Agency and Expert Group on SDGs, and those may help further specify objective investigative efforts (IAEG-SDGs 2016).

Because the UN SDGs are already so widely applied by responsible investors, significant work has been done to account for the contribution – or subtraction – of AI to sustainability. One publication reports:

> In SDG 1 (No poverty), SDG 4 (Quality education), SDG 6 (Clean water and sanitation), SDG 7 (Affordable and clean energy), and SDG 11 (Sustainable cities), AI may act as an enabler for all the targets by supporting the provisioning of food, health, water, and energy services to the population. AI can enable smart and low-carbon cities encompassing a range of interconnected technologies such as electrical autonomous vehicles and smart appliances that can enable demand response in the electricity sector with benefits across SDGs 7, 11, and 13 on climate action (Vinuesa et al. 2020: 2).

Technology can also diminish sustainability:

> AI may also lead to additional qualification requirements for any job, consequently increasing the inherent inequalities and acting as an inhibitor towards the achievement of this target (Vinuesa et al. 2020: 2-3).

The next step is to repeat the industry analysis, but as applied to particular AI-intensive companies. For example, the firm AgrilogicAI (later rebranded as Dagan Tech) uses machine learning methods to identify high-yielding soybean variants by analyzing data from remote sensing and soil features. According to their report:

> Collectively, our models identified fifteen elite varieties from 21 predictive variables to forecast soybean yields in 2015 at 58 test locations. This method can boost commercial soy yields by about 5% and shorten the time for commercial variant development (Aviv 2018).

So, the technology helps maximize yield for specific soil conditions, and speeds crop optimization, which serves SDGs 1 and 2 (Poverty, Hunger), as well as 15, which seeks to protect, restore and promote sustainable use of terrestrial ecosystems. Similar points could be made about the company's *Farm360AI* platform which predicts corn and soybean yields from satellite imagery and weather data. Currently, the business is privately held, but it

nevertheless exemplifies a high score in the sustainability metric of the social wellbeing principle.

In summary, sustainability as an ethical investment criterion measures social wellbeing as enduring, and it can be scored as AI-intensive companies' addition to, or subtraction from progress toward the 17 UN Sustainable Development Goals. Data corresponding with major companies' performance – as well as ratings providers' results – are abundant, and available to investors. Pricewaterhouse Coopers's 2019 SDG Challenge investigates over 1,000 company reports on their engagement with sustainability (Scott and McGill 2019). In 2020, S&P Global analyzed 150 categories aligned with the SDGs across 3,500 companies representing 85% of global market capitalization (Trucost 2020). Bloomberg reports that at least a dozen major third-party enterprises provide independent ratings of companies' SDG sustainability performance, including Sustainalytics, MSCI, Moody's, and Fitch Ratings (Poh 2019).

Figure 7

AI Ethics of Social Wellbeing	
Sustainability: *Does the AI promote enduring social wellbeing?*	**Negative** • **Slows progress toward the 17 UN Sustainable Development Goals.** **Positive** • **Speeds progress toward the 17 UN Sustainable Development Goals.**

Technical Trustworthiness Metrics: Performance, Safety, Accountability

Performance measures the accuracy of AI outputs, along with the efficiency of their production. In human experience, performance may be evaluated in terms of personalization and convenience.

At a 2020 professional AI conference, a Netflix machine learning research scientist was asked, "How is Netflix using AI for a positive impact?" He responded:

> We try to build models for recommendations that maximize Netflix member's enjoyment of the selected item while minimizing the time it takes to find them. Enjoyment integrated over time i.e. goodness of the item and the length of view, interaction cost integrated over time i.e. time it takes the member to find something to play, are some of the factors we consider while building our ML/AI models for a positive impact on our 100M+ members (Deoras 2020).

For distinct AI companies and functions, the meaning of personalized quality and convenience will shift, but the Netflix metrics – enjoyment integrated over time, interaction cost integrated over time – double as objective ethical scores. In a sense, performance is the easiest ethical investing criterion: the more accurate and efficient the AI is technically, the better it is ethically.

Another way to score performance is in relative terms, as compared with other AIs. If you could use only Google or Bing for a year, which would you choose? Market share may provide a simple and revealing answer to this question about measuring personalized convenience.

A similarly relative evaluation could be performed with AI set against human providers. In 2019 Google computer scientist Geoffrey Hinton tweeted a memorable thought experiment that circulated widely:

> Suppose you have cancer and you have to choose between a black box AI surgeon that cannot explain how it works but has a 90% cure rate and a human surgeon with an 80% cure rate? (Hinton 2020.)

The answer – perhaps as provided by a focus group or, more pointedly, by actual patients in a hospital – may produce a useful and double evaluation of AI accuracy and efficiency as they relate to the larger principle of technical trustworthiness. First, the raw numbers could be tested: Does the AI really outperform the human? Second, the human weight of the comparison could be evaluated: How large must the outperformance be for patients to opt for the AI?

It is now documented that in certain medical fields AI *does* outperform humans not just in diagnosis, but in what can be processed on the way to diagnosing (Grote 2019). Cancer screenings, for example, test human doctors in two ways: sensitivity to anomalies (how well they see), and rapidity of scan readings (how much they see). The latter is increasingly significant because new cancer detection technology seeks to increase sensitivity by multiplying images (Conant et al. 2019). There is no upper limit, millions of scans could be sliced from any one patient. For human doctors, however, most would be superfluous since there are not enough hours in the day to examine them all. AI does not have that problem: the machine could theoretically scan the images as rapidly as they are produced. It follows that a true performance rating will not only account for how well a task is accomplished (image analysis), but what tasks become possible (rapidly analyzing streams of images) when AI is doing the performing.

Performance as an indicator of human impact investing is indispensable and influential. Without it there is no investing to impact: performance is the condition of the possibility of doing AI ethics. And, as performance increases, so too does use and correspondingly the human effect: performance is an AI ethics multiplier. As a criterion of technical trustworthiness, it is measured as accuracy and efficiency, or as personalization and convenience.

Figure 8

AI Ethics of Technical Trustworthiness	
Performance: ***How accurate and efficient is the machine, how personalized and convenient the output?***	**Negative** • **Accuracy, efficiency, personalization, convenience inferior to competitors in category, or human-based alternatives.** **Positive:** • **Accuracy, efficiency, personalization, convenience superior to competitors in category *and* human-based alternatives.**

Safety as a criterion of technical trustworthiness asks whether an AI is resilient, and empowered with fallbacks to mitigate failures.

In 2016, a Tesla crashed into a truck. According to Tesla:

> The Model S was on a divided highway with Autopilot engaged when a tractor trailer drove across the highway perpendicular to the car. Neither Autopilot nor the driver noticed the white side of the tractor trailer against a brightly lit sky, so the brake was not applied. The high ride height of the trailer combined with its positioning across the road and the extremely rare circumstances of the impact caused the Model S to pass under the trailer, with the bottom of the trailer impacting the windshield of the Model S. (Tesla Team. 2016.)

This horror movie accident represents a particular AI fear: a machine capable of calculating pi to 31 trillion digits (Porter 2019) cannot figure out to stop when a truck crosses in front. The power juxtaposed with the debility seems ludicrous, as though the machine completely lacks common sense which, in fact, is precisely what it does lack (De Freitas 2020: 8).

For human users, one grievous effect of the debility is no safe moment. As with any mechanism, AIs come with knowable risks and dangers, but it is beyond that, in the region of unknowable risks – especially those seemingly easily avoidable, even for children – that human trust in AI destabilizes.

The Stanford Center for AI Safety reports that "machine-learned systems are highly complex, and that humans can barely parse their mathematical formulas" (Barrett 2019: 1). So, part of the reason an AI can shockingly and unexpectedly fail is the convoluted nature: when there is no practical way to follow how the machine is working, it becomes impossible to predict what sudden catastrophe might come next.

At least theoretically, the complexity is resolvable: an infinite human intellect could presumably keep pace with the deepest neural network. That only reveals a deeper problem, however. Machines and humans create knowledge differently. AI filters for correspondences, while humans impose linear causality onto raw perceptions (Kant 1997: A91/B124). This difference –

correspondence versus causality – at the origin of knowledge itself means that machine learning *cannot* be understood, even by infinitely quick human thinking. The true divergence, in other words, does not concern velocity or power of reasoning, instead, it is about what the verb *to reason* means, and that renders everything coming afterward irreconcilable. So, the way AI produces knowledge is inhuman, which does not falsify the knowledge, but it does preclude comprehensive human knowledge about the knowledge.

Because decisions guided by pure correspondence will always be prone to humanly incomprehensible failures, scoring an AI company for safety becomes disorienting. It is not just that perfect confidence is impossible, but also that there is no way to limit the scenes of peril. Consequently, safety can only be conceived as a process instead of a goal: improvements may be marked by remedying encountered problems like the Tesla failing to detect the truck, but since the remaining risk cannot be calculated, there is no way to know that a safety goal has been reached, or even how close we are to it.

In their paper *A Safety Standard Approach for Fully Autonomous Vehicles* the authors write that, "it is important to address known safety issues before exposing testers and the public to undue safety risk," and add:

> Rather than adopting a fiction that mere conformance to a standard at deployment results in flawless risk mitigation, it is important to continually evaluate and improve the residual risk present in the system. Honest self-assessment and iteration over the system development and deployment lifecycle is vitally important to mature the safety case (Koopman et al. 2019: 6).

This is a roundabout way of saying that we should make driverless cars as safe as reasonably possible at the start, and then when accidents occur, learn why as best we can and redesign the AI to avoid future recurrences. Users, that means are irretrievably crash test dummies.

One response to this safety challenge is human oversight. A designer monitoring the AI ("Human on the loop") or a deployment supervisor accompanying the AI ("Human in control") holds power to adjust what is happening, and therefore merits responsibility for the machine's actions,

especially those easily countermanded by human common sense. Tesla tapped into this AI safety strategy when responding to the truck crash. After explaining what happened out on the road, and expressing condolences, the company curtly added:

> When drivers activate Autopilot, the acknowledgment box explains that Autopilot "is an assist feature that requires you to keep your hands on the steering wheel at all times," and that "you need to maintain control and responsibility for your vehicle" while using it (Tesla Team. 2016).

The message is that human users are empowered to override AI decisions and so shave off at least those dangers obvious to us and invisible to machines. On the other hand, if users need to be driving along, what is the point of Autopilot?

Ultimately, safety as an AI human impact performance indicator needs to be rendered calculable and meaningful. The E.C. *Guidelines to Trustworthy AI* lists considerations that analysts could convert into scoreable categories. They include:

- Protections against hacking.
- Accounting for unintended uses.
- Fallback systems that ask for human operation in the face of irresolvable problems (AIHLEG: 16-17).

Another measuring possibility uses resource allocation as a safety proxy: the more money and expertise a company dedicates to ensuring its AI mistakes are rapidly and well corrected, the higher its score. Of course, the importance of safety itself depends on the magnitude of risk posed by a system's capabilities (AIHLEG: 17): driverless cars and autonomous floor cleaners present distinct dangers and require different investments to qualify as safe.

Empirical safety results may also be measured either across the industry, or comparatively between AIs and humans. A safety score for Tesla may be initially calculated by weighing deaths per mile driven (or collisions per mile driven, or a similar anomaly) against those attributable to other autonomous vehicle companies. The same strategy could be applied between Tesla and

human drivers. Either way, there are no guarantees. There are not even confident probabilities of safety since we can never know the full extent of the risk of what might go wrong.

Figure 9

AI Ethics of Technical Trustworthiness	
Safety: ***Is the AI resilient, and empowered with fallbacks to mitigate failures?***	**Negative:** • **Inadequate protections against hacking and unintended uses** • **Inconsistent fallback mechanisms** • **Resources dedicated to continuous safety review incommensurate with risk** • **AI statistically less safe than competitors in category *or* human-based alternatives** **Positive:** • **Robust protections against hacking and unintended uses** • **Efficient fallback mechanisms** • **Resources dedicated to continuous safety review commensurate with risk** • **AI statistically safer than competitors in category *and* human-based alternatives.**

Accountability is the third technical trustworthiness indicator. When an AI goes wrong – and when it goes right – professional accountability measures how well responsibility is assigned.

One way to train an autonomous vehicle is through demonstration. A human driver takes the lead by operating a camera-outfitted car, and as kilometers and data accumulate, the AI is increasingly able to make decisions by imitating those observed on the road (Kebria 2020). Potentially, human owners could train their vehicles with traits of their own driving. Distance between cars while cruising, acceleration rate, breaking abruptness, turning radiuses, all that could be personalized. Later, and with autopilot engaged, the car crashes. Who is to blame? The owner? The AI? The AI designer?

Determining responsibility starts with explainability: explaining *why* an AI produced a specific output or decision is required to locate the source of a fault. Without that, there is nowhere for the process to begin. In the 2016 Tesla crash, a report from the European Parliament found that the Tesla AI mistook the white tractor trailer crossing in front for the sky, and did not even slow down (Panel 2020: 34). That made for a frightful accident, but a comprehensible AI event: because it was possible to grasp the algorithmic decision in human terms, an investigation into responsibility became possible. The potential conclusions are multiple. The original AI training may have been insufficient, perhaps the machine should have been fed more images of trucks from the side view to refine its recognition. Or, the car camera may have been imperfectly filtered for solar glare. Possibly the human driver should have been more attentive. There are further options, and deciding between them belongs to a discussion that is well rehearsed in business ethics and liability law (Roe 2019; Casy 2019). The central point, however, is that accountability requires explainability: decisions made by an AI system need to be understandable and traceable by human beings (AIHLEG: 18).

There are at least four challenges to the explainability requirement. First, as noted in the Safety section above, the AI may not be a black box so much as a black hole in the sense that its workings cannot be rendered sensible to the human mind. So, investigators may be able to determine that the truck was digitally recognized as the sky, but unable to comprehend how or why that misinterpretation of the data arose. And, as AI-produced knowledge advances, it becomes increasingly evident that human understanding may not be fitted to the machine's algorithmic methods (Zerilli 2019: 670). In this case, a lesser standard may be applied, possibly interpretability (Gall 2018), which is about predicting what will happen instead of why: it is foreseeing what output will follow from an input, as opposed to following along to determine exactly how the output gets generated.

Besides explainability and interpretability, other terms commonly employed to discuss the backward engineering of AI decisions include transparency and auditability (AIHLEG: 13). For any ethical evaluation of AI accountability to

succeed, technical experts will be required to help navigate the computer science terrain and discover the boundaries of humanly understanding a machine's operation.

A further complication of the explainability demand is that it may harm AI performance (Whittlestone et al 2019: 20). If humans cannot keep up, and explainability is a priority, then the machine can be slowed down: Gaussian processes, neural networks and random forests can be replaced by linear regression or a single decision tree. These changes represent, in effect, an exchange of computing power for a better view of the computations. The choice behind it – knowing, or knowing why there is knowing – may be answered one way or the other, but in terms of ethical investing, a positive explainability score may derive from the willingness to sacrifice at least some performance.

The third problem with the explainability demand is that the algorithms may be gathered as a trade secret, creating a dilemma between economic and ethical viability.

Finally, there are many cases where explainability is gratuitous because accountability is superfluous. When an AI is recommending neckties or jazz playlists, harm done by errors – and consequently the ethical need for correctly assigning responsibility – is vanishingly small.

When accountability is not superfluous, explainability must be scored positively or negatively. There are two broad approaches. One gauges transparency directly: in objective terms, how much of an AI's inner workings can be traced, and how does that understanding compare with other AIs performing similar tasks? Further, how does it compare with similarly tasked *humans?* If a manager or hiring supervisor or doctor relies on gut feelings to make decisions – if the comparable human thought process is not explainable – it is reasonable to enforce only minimal transparency demands on AI functioning (Alufaisan et al. 2020: 3).

Another approach to explainability measuring is through tradeoffs: how much is a company willing to sacrifice in terms of performance, or profit to reveal how the AI processes from input to output?

Accountability begins with explainability and the assignment of responsibility. Next comes *redress*, which is the ability to respond to – or recover compensation for – actions stemming from the AI function. Legal studies influence significantly here, including the doctrine of the learned intermediary (Harned et al. 2019) which generally holds in the medical field that when a machine error causes harm, the human operator is responsible. If an AI cancer diagnosis proves erroneous, it is the doctor who signed the finding, not the machine that would be the target of a lawsuit (Sullivan 2019).

There is a structural problem, however. A central AI benefit lies in an increasing ability to function without human oversight: the reason Tesla is developing autopiloting cars is not so that drivers can ride along gripping the wheel, it is to allow a nap on the way home from work. So, the further a machine learning platform advances, the more it may extend from any learned intermediary, and that means the better the AI, the harder it is to pin blame on a human overseer.

One solution is to blame the AI directly by assigning legal personhood to it, as is done for corporations (Hildebrandt 2019).

Another solution is redress-by-design (Quintarelli 2019), which is the engineering strategy of formulating AIs so that harms can be not only identified and corrected rapidly, but also so that outputs can be *contested* (Ploug and Holm 2020; GDPR: Article 22). For example, AI is increasingly employed to make lending decisions (Verma and Rubin 2018) because loan distribution can be reduced to predictive analytics estimating the risk of default. When a loan is denied, redress-by-design may help applicants understand what specific piece of data led to their rejection, and enable the opportunity to object effectively.

More broadly, the E.C. *Guidelines for Trustworthy AI* establishes elements for adequate redress in AI systems (AIHLEG 2019: 31). There should be:

- a way for users to contest decisions
- a way for users to make a claim for harm done
- available information about how to make the claim
- available information about the circumstances that may occasion such a claim

In the end, accountability as an ethical investment performance indicator measures how well responsibility can be assigned for what an AI does. It encompasses – and can be scored as – explainability and redress.

Figure 10

AI Ethics of Technical Trustworthiness	
Accountability: ***How well can responsibility be assigned for what an AI does?***	**Negative** • **Limited explainability and transparency** • **Few or obscure mechanisms for contestation and redress** **Positive** • **Significant explainability or transparency** • **Ample contestation and redress opportunities**

How are the Categories Scored?

Following Vinuesa (Vinuesa et al. 2020), a three-point metric may be employed to score ethical performance indicators in AI intensive companies. A score of 2 corresponds with a positive evaluation, 1 corresponds with neutral or not material, and 0 corresponds with inadequacy. The scores convert into objective investment guidance, both individually and as a summed total.

Investors who are particularly interested in privacy, for example, or safety, may choose to highlight those metrics in their analysis of investment opportunities. Others may widen the humanitarian vision to include the full

range of AI ethics concerns, and so focus on the overall impact score derived from a company.

The scoring rubric is itemized below. Because personal freedom is the orienting metric of responsible AI investing, it is double weighted in the aggregating formula.

- Personal Freedom
    - Autonomy (Score 0 - 2)
    - Dignity (Score 0 - 2)
    - Privacy (Score 0 - 2)
- Social Wellbeing
    - Fairness (Score 0 - 2)
    - Solidarity (Score 0 - 2)
    - Sustainability (Score 0 - 2)
- Technical Trustworthiness
    - Performance (Score 0 - 2)
    - Safety (Score 0 - 2)
    - Accountability (Score 0 - 2)
- AI Human Impact Score
    - Total PF*2 + SW + TT = (0 – 24) / 2.4 = Net Score on 10 scale

Conclusion

ESG investing, along with ethical, responsible, sustainable, and impact variations, is commanding increasing assets (GSIA 2019), and widening in appeal. The appeal initiated with investors promoting social change (Tett 2019), and now includes executives seeking to limit reputational and financial risk (Fink 2020). Meanwhile, AI is consuming a larger share of the global economy: one calculation has the industry utilizing about 1% of the world's electricity in 2018, but reaching 20% by 2030 (Vinuesa 2020).

While ESG investing and AI are growing, they are also growing apart. ESG was forged in the industrial economy amid competition between social collectives and conflicts over the exploitation of natural resources (MSCI 2020). The fundamental tension of the AI economy is different, it is *personalized*. What most matters today is the question about individuals and their own data: Is gathered personal information processed to invigorate self-determination and expand opportunities, or does it narrow possible human experiences?

For the question to be addressed by ethical finance, a new model is needed for evaluating companies that operate with AI at their core. The model should identify investments that vitalize personal freedom, while also supporting social wellbeing and fortifying technical trustworthiness. The model should be intellectually robust, manageable for analysts, useful for portfolio managers, and credible for investors. AI Human Impact is one possibility.

References

(AIHLEG) Artificial Intelligence High Level Expert Group. (2019). *Ethics Guidelines for Trustworthy AI*, European Commission. Accessed 2 July 2020: https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

Allen, Mike. (2017). Sean Parker unloads on Facebook: ‘God only knows what it's doing to our children's brains,’ *Axios*, Nov 9. Accessed 2 July 2020: https://www.axios.com/sean-parker-unloads-on-facebook-2508036343.html

Alufaisan, Yasmeen; Marusich, Laura; Bakdash, Jonathan; Zhou, Yan and Murat Kantarcioglu. (2020). Does Explainable Artificial Intelligence Improve Human Decision-Making? arxiv.org. arXiv:2006.11194v1. Accessed 7 July 2020: https://arxiv.org/abs/2006.11194

Angwin, Julia; Larson, Jeff; Mattu, Surya and Lauren Kirchner. (2016). Machine Bias. *ProPublica*. May 23. Accessed 2 July 2020:

https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Antoncic, Madelyn. (2020). Uncovering hidden signals for sustainable investing using Big Data: Artificial intelligence, machine learning and natural language processing, *Journal of Risk Management in Financial Institutions*, Volume 13, Number 2, Spring, pp. 106-113(8).

Aristotle. (1934). Nicomachean Ethics, *Aristotle in 23 Volumes, Vol. 19*, trans. H. Rackham. Cambridge, MA: Harvard University Press.

Aviv, Tzvi. (2018). Ensemble of Cubist models for soy yield prediction using soil features and remote sensing variables. KDD2017 Conference. Accessed 2 July 2020: https://www.youtube.com/watch?v=RXgASdsrxl0

Barocas S. and A. Selbst. (2016). Big Data's Disparate Impact. *California Law Review*, 104(3): 671–732. DOI: 10.2139/ssrn.2477899

Barrett, Clark; Dill, David; Kochenderfer, Mykel and Dorsa Sadig. (2019). White Paper. Stanford Center for AI Safety. Accessed 2 July 2020: http://aisafety.stanford.edu/whitepaper.pdf

Berinato, Scott. (2014). OkCupid's Co-Founder Probably Wouldn't Agree to the Experiments OkCupid Runs on Its Users, *Harvard Business Review Blog*, July 29. Accessed 2 July 2020: https://hbr.org/2014/07/okcupids-co-founder-probably-wouldnt-agree-to-the-experiments-okcupid-runs-on-its-users

Bobrowski, D. and H. Joshi. (2019). Unmasking A.I.'s bias in healthcare: The need for diverse data, *University of Toronto Medical Journal*: 96.

Borenstein, Jason and Ronald Arkin. (2017). Nudging for good: robots and the ethical appropriateness of nurturing empathy and charitable behavior. *AI and Society:* 32, 4 (November), 499–507. DOI: https://doi.org/10.1007/s00146-016-0684-1

Bozdag, Engin and Jeroen van den Hoven. (2015). Breaking the Filter Bubble: Democracy and Design, *Ethics and Information Technology*: 17, 4 (01 Dec),249–265.

Reiser, Dana and Anne Tucker. (2019). Buyer Beware: Variation and Opacity in ESG and ESG Index Funds (Dec 1). Accessed 2 July 2020: https://ssrn.com/abstract=3440768.

Caliskan, Aylin; Bryson, Joanna and Arvind Narayanan. (2017). Semantics Derived Automatically from Language Corpora Contain Human-Like Biases, *Science*, 14 Apr: Vol. 356, Issue 6334, pp. 183-186. DOI: 10.1126/science.aal4230.

Casey, Bryan. (2019). Robot Ipsa Loquitur, *Georgetown Law Journal*, January 20. DOI: http://dx.doi.org/10.2139/ssrn.3327673.

Cavoukian, Ann. (2011). *Privacy by Design: The 7 Foundational Principles*. Information & Privacy Commissioner Ontario, Canada. Accessed 2 July 2020: https://iapp.org/media/pdf/resource_center/pbd_implement_7found_principles.pdf.

(CCPA) *California Consumer Privacy Act Fact Sheet*. (2019). California Department of Justice, Office of the Attorney General. Accessed 2 July 2020: https://www.oag.ca.gov/system/files/attachments/press_releases/CCPA%20Fact%20Sheet%20%2800000002%29.pdf.

Celis, Elisa; Kapoor, Sayash; Salehi, Farnood and Nisheeth Vishnoi. (2019). Controlling Polarization in Personalization: An Algorithmic Framework, in *Proceedings of the Conference on Fairness, Accountability, and Transparency* (FAT* ’19). Association for Computing Machinery, New York, NY, USA, 160–169. DOI: https://doi.org/10.1145/3287560.3287601

Chan, S.; Godwin, H. and A Gonzalez. (2017). Review of Use and Integration of Mobile Apps into Psychiatric Treatments. *Current Psychiatry Reports,* 19, 96. DOI: https://doi.org/10.1007/s11920-017-0848-9

Char, D.; Shah, N. and D. Magnus. (2018). Implementing Machine Learning in Health Care — Addressing Ethical Challenges. New, *England Journal of Medicine*, 378, pp. 981-983.

Clement-Jones, Tim and Luciano Floridi. (2019) The Five Principles Key to Any Ethical Framework for AI. *New Statesman*, 20 March. Accessed 2 July 2020: https://tech.newstatesman.com/policy/ai-ethics-framework

Conant, Emily; Toledano, Alicia; Periaswamy, Senthil; Fotin, Sergei; Go, Jonathan; Boatsman, Justin and Jeffrey Hoffmeister. (2019). Improving Accuracy and Efficiency with Concurrent Use of Artificial Intelligence for Digital Breast Tomosynthesis, *Radiology: Artificial Intelligence*; 1 (4): e180096. DOI: 10.1148/ryai.2019180096.

De Freitas, Julian, Andrea Censi, Luigi Di Lillo, Sam E. Anthony, and Emilio Frazzoli. (2020). "From Driverless Dilemmas to More Practical Ethics Tests for Autonomous Vehicles." PsyArXiv. May 10. DOI:10.31234/osf.io/ypbve

(DEC) Data Ethics Commission of the Federal Government Federal Ministry of the Interior, Building and Community. (2019). Opinion of the Data Ethics Commission. Accessed 2 July 2020: https://www.bmjv.de/SharedDocs/Downloads/DE/Themen/Fokusthemen/Gutachten_DEK_EN.pdf? blob=publicationFile&v=2

Deleuze, Gilles. (1992). Postscript on the Societies of Control. *October* 59 (Winter): 3-7. Original publication: Deleuze, Gilles. Post-scriptum sur les sociétés de contrôle. *L 'autre journal*, n°1, mai 1990. Accessed 2 July 2020: https://infokiosques.net/lire.php?id_article=214

Deoras, Anoop. (2020). How Netflix uses AI to Predict Your Next Series Binge. *ReWork Blog*. Accessed 2 July 2020: https://blog.re-work.co/how-does-netflix-use-ai-to-predict-your-next-series-binge/

Eccles, Robert G. and Judith Stroehle. (2018). Exploring Social Origins in the Construction of ESG Measures (July 12). Accessed 2 July 2020: https://ssrn.com/abstract=3212685 or http://dx.doi.org/10.2139/ssrn.3212685

Fink, Laurence. (2020). A Fundamental Reshaping of Finance. BlackRock. Accessed 2 July 2020: https://www.blackrock.com/corporate/investor-relations/larry-fink-ceo-letter

Fjeld, Jessica, Achten, Nele, Hilligoss, Hannah, Nagy, Adam and Madhulika Srikumar. (2020). *Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI* (January 15). Berkman Klein Center Research Publication No. 2020-1. DOI: http://dx.doi.org/10.2139/ssrn.3518482

Floridi, Luciano and Josh Cowls. (2019). A Unified Framework of Five Principles for AI in Society, *Harvard Data Science Review,* Jul 01. DOI: https://doi.org/10.1162/99608f92.8cd550d1

Gall, Richard. (2018). Machine Learning Explainability vs Interpretability: Two concepts that could help restore trust in AI. *KDnuggets.com*. Accessed 2 July 2020: https://www.kdnuggets.com/2018/12/machine-learning-explainability-interpretability-ai.html

(GDPR) General Data Protection Regulation. (2016). General Data Protection Regulation, European Parliament and of the Council of the European Union. 27 April. Accessed 2 July 2020: https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32016R0679&from=EN

Gianfrancesco, M.; Tamang, S.; Yazdany, J. and G. Schmajuk. (2018). Potential Biases in Machine Learning Algorithms Using Electronic Health Record Data, *JAMA Intern. Med*., 178, pp. 1544-1547.

Goodman, S.; Goel, S. and M.R. Cullen. (2018). Machine Learning, Health Disparities, and Causal Reasoning, *Annals of Internal Medicine*, 169, p.883.

Griffiths, Robert. (2016). New Zealand Passport Robot Thinks This Asian Man's Eyes Are -Closed, *CNN.com*, December 9. Accessed 2 July 2020: https://www.cnn.com/2016/12/07/asia/new-zealand-passport-robot-asian-trnd/

Grote, Thomas and P. Berens. (2019). On the ethics of algorithmic decision-making in healthcare, *Journal of Medical Ethics*; 0:1–7. DOI: 10.1136/medethics-2019-105586.

(GSIA) Global Sustainable Investment Alliance. (2018). *Global Sustainable Investment Review*. Accessed 2 July 2020: http://www.gsi-alliance.org/wp-content/uploads/2019/06/GSIR_Review2018F.pdf

Hagendorff, T. (2020). The Ethics of AI Ethics: An Evaluation of Guidelines, *Minds and Machines* 30, 99–120. DOI: https://doi.org/10.1007/s11023-020-09517-8

Hardt, Moritz; Price, Eric and Nathan Srebro. (2016). Equality of Opportunity in Supervised Learning. *30th Conference on Neural Information Processing Systems* (NIPS 2016), Barcelona, Spain. Accessed 2 July 2020: https://papers.nips.cc/paper/6374-equality-of-opportunity-in-supervised-learning.pdf

Harned, Zach; Lungren, Matthew and Pranav Rajpurkar. (2019). Comment, Machine Vision, Medical AI, and Malpractice, *Harvard Journal of Law & Technology Digest*. Accessed 2 July 2020: https://jolt.law.harvard.edu/digest/machine-vision-medical-ai-and-malpractice.

Hildebrandt, M. (2019). *Legal Personhood for AI*. Law for Computer Scientists and Other Folk.

Hill, Kashmir. (2017). How Facebook Outs Sex Workers, *Gizmodo*. Accessed 2 July 2020: https://gizmodo.com/how-facebook-outs-sex-workers-1818861596. Accessed 2 July 2020.

Hinton, Geoffrey. (2020). Twitter.com. 3:37 PM · Feb 20. https://twitter.com/geoffreyhinton/status/1230592238490615816

(IAEG-SDGs) Inter-Agency and Expert Group on Sustainable Development Goal Indicators. (2016). Final list of proposed Sustainable Development Goal indicators, (E/CN.3/2016/2/Rev.1) 2/25 Annex IV. Accessed 2 July 2020: https://sustainabledevelopment.un.org/content/documents/11803Official-List-of-Proposed-SDG-Indicators.pdf

(IEEE) IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. (2019). *Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems*, First Edition. IEEE. Accessed 2 July 2020: https://standards.ieee.org/content/ieee-standards/en/industry-connections/ec/autonomous-systems.html.

Jobin, Anna; Ienca, Marcello and Effy Vayena. (2019). Artificial Intelligence: the global landscape of ethics guidelines, *Health Ethics & Policy Lab*, ETH Zurich. Accessed 2 July 2020: https://arxiv.org/ftp/arxiv/papers/1906/1906.11668.pdf

Kamehkhosh, I., Bonnin, G. and D. Jannach. (2020). Effects of recommendations on the playlist creation behavior of users, *User Model User-Adap Inter:* 30, 285–322. DOI: https://doi.org/10.1007/s11257-019-09237-4

Kant, Immanuel. (1996). *Groundwork of the Metaphysics of Morals* in *Immanuel Kant: Practical Philosophy*, trans. Mary Gregor. (Cambridge, England: Cambridge University Press).

Kant, Immanuel. (1997). *Critique of Pure Reason*, trans. Paul Guyer and Allen W. Wood (Cambridge: Cambridge University Press).

Kirkpatrick, D. (2011). *The Facebook Effect*. New York: Simon & Schuster.

Kebria, P.; Khosravi, A.; Salaken, S. and S. Nahavandi. (2020). Deep imitation learning for autonomous vehicles based on convolutional neural networks, *IEEE/CAA, Journal of Automatica Sinica*, vol. 7, no. 1, pp. 82–95, Jan.

Kodiyan, Akhil Alfonse. (2019). An overview of ethical issues in using AI systems in hiring with a case study of Amazon's AI based hiring tool, *Research Gate*. November 12. https://www.researchgate.net/profile/Akhil_Kodiyan/publication/337331539_An_overview_of_ethical_issues_in_using_AI_systems_in_hiring_with_a_case_study_of_Amazon's_AI_based_hiring_tool/links/5dd2aa8d4585156b351d330a/An-overview-of-ethical-issues-in-using-AI-systems-in-hiring-with-a-case-study-of-Amazons-AI-based-hiring-tool.pdf

Koopman, Philip; Ferrell, Uma; Fratrik, Frank and Michael Wagner. (2019). 'A Safety Standard Approach for Fully Autonomous Vehicles' in *Computer Safety, Reliability, and Security.* (Switzerland: Springer Nature AG). DOI: 10.1007/978-3-030-26250-1_26.

Lam, Onyi; Wojcik, Stefan; Broderick, Brian and Adam Hughes. (2018). *Gender and Jobs in Online Image Searches*, Pew Research Center, December 17. Accessed 2 July 2020: https://www.pewsocialtrends.org/2018/12/17/gender-and-jobs-in-online-image-searches/

Mason, C. and B. Doherty. (2016) *A Fair Trade-off? Paradoxes in the Governance of Fair-trade Social Enterprises*, Journal of Business Ethics 136, 451–469. DOI: https://doi.org/10.1007/s10551-014-2511-2

Microsoft. (2018). Inclusive Design Toolkit. Accessed 2 July 2020: https://www.microsoft.com/design/inclusive/

Mittelstadt, Brent. (2019). Principles alone cannot guarantee ethical AI, *Nature Machine Intelligence*. DOI: 10.1038/s42256-019-0114-4.

Mohamadi, A., Heidarizadi, Z. and H. Nourollahi. (2016). Assessing the desertification trend using neural network classification and object-oriented techniques. *J.Fac. Istanb. Univ*. 66, 683–690.

MSCI. (2014). *ESG Issue Report: Privacy and Data Security - Exploring the Data Value Chain*. July. Accessed 2 July 2020: https://www.msci.com/documents/10199/0edc73b8-daf7-447d-9ff8-f487c740c560

MSCI. (2019). *MSCI ESG Ratings Methodology*. Accessed 2 July 2020: https://www.msci.com/documents/1296102/14524248/MSCI+ESG+Ratings+Methodology+-+Exec+Summary+2019.pdf/2dfcaeee-2c70-d10b-69c8-3058b14109e3?t=1571404887226

MSCI. (2019b). *ESG Sustainable Impact Metrics: Incorporating Sustainable Impact in Your Investment Process*, MSCI ESG Research LLC. Accessed 2 July 2020: https://www.msci.com/documents/1296102/16472518/ESG_ImpactMetrics-cfs-en.pdf/7a03ddab-46fd-cef7-5211-c07ab992d17b?t=1574374138495

MSCI. (2020). *The Evolution of ESG Investing*. Accessed 2 July 2020: https://www.msci.com/what-is-esg

Narayanan, Arvind. (2018). *21 Definitions of Fairness* at *ACM FAT* '18 Conference*, New York, NY, USA. Accessed 2 July 2020: https://www.youtube.com/watch?v=jIXIuYdnyyk.

Narayanan, Arvind; Mathur, Arunesh; Chetty, Marshini and Mihir Kshirsagar. (2020) *Dark Patterns: Past, Present, and Future*. Acmqueue: march-april. Accessed 2 July 2020: https://dl.acm.org/doi/pdf/10.1145/3400899.3400901

Nedelkoska, Ljubica and Glenda Quintini. (2018), 'Automation, skills use and training,' OECD Social, Employment and Migration Working Papers, No. 202, OECD Publishing, Paris. DOI: https://doi.org/10.1787/1815199X

Nolan, Christopher; Lawyer, Glenn and Ryan Marshall Dodd. (2019). Cybersecurity: today's most pressing governance issue, *Journal of Cyber Policy*, 4:3, 425-441, DOI: 10.1080/23738871.2019.1673458

Noor, Poppy. (2002). Can we trust AI not to further embed racial bias and prejudice? *BMJ* 368:m363. DOI: https://doi.org/10.1136/bmj.m363

(Panel) Panel for the Future of Science and Technology. (2020). The ethics of artificial intelligence: Issues and initiatives. European Parliament. March. Accessed 2 July 2020: https://www.europarl.europa.eu/RegData/etudes/STUD/2020/634452/EPRS_STU(2020)634452_EN.pdf

Pereira, J., Díaz, Ó. (2019). Using Health Chatbots for Behavior Change: A Mapping Study. *J Med Syst* 43**,** 135. DOI: https://doi.org/10.1007/s10916-019-1237-1

Ploug T. and S. Holm. (2020). The four dimensions of contestable AI diagnostics - A patient-centric approach to explainable AI, *Artificial Intelligence in Medicine:* IF 3.574. DOI: https://doi.org/10.1016/j.artmed.2020.101901

Poh, Jacqueline. 2019. *Conflicting ESG Ratings Are Confusing Sustainable Investors*. Bloomberg. December 11. Accessed 2 July 2020: https://www.bloomberg.com/news/articles/2019-12-11/conflicting-esg-ratings-are-confusing-sustainable-investors

Porter, Jon. 2019. Google employee calculates pi to record 31 trillion digits. *The Verge*. Mar 14. Accessed 2 July 2020: https://www.theverge.com/2019/3/14/18265358/pi-calculation-record-31-trillion-google

Press release. (2019). Fight for the Future launches major new campaign calling for a Federal ban on facial recognition surveillance. FightforthrFuture.org. July 9. Accessed 2 July 2020: https://www.fightforthefuture.org/news/2019-07-09-fight-for-the-future-launches-major-new-campaign/

Quintarelli, Stefano. 2019. We need "redress by design" for AI systems, Quintarelli Blog. Accessed 2 July 2020: https://blog.quintarelli.it/2019/04/we-need-redress-by-design-for-ai-systems.html

Rawls, John. (1971). *A Theory of Justice*. Boston: Harvard University Press. ISBN 0674000781. OCLC 41266156.

Remia, Mahajan. (2015). Addicted to Facebook: Role of Narcissism and Self-Esteem, *Zenith: International Journal of Multidisciplinary Research*, Volume: 5, Issue: 10, pp. 159 – 165.

Roe, Madeline. (2019). Who's Driving That Car?: An Analysis of Regulatory and Potential Liability Frameworks for Driverless Cars, 60 *B.C.L. Rev. 315*. Accessed 2 July 2020: https://lawdigitalcommons.bc.edu/bclr/vol60/iss1/7

Rudder, Christian. (2014). We Experiment on Human Beings! *OKCupid Blog*, July 28. Accessed 2 July 2020: http://web.archive.org/web/20140728192045/http://blog.okcupid.com/index.php/we-experiment-on-human-beings/#expand

Sakulkar, Pranav and Bhaskar Krishnamachari. (2016). Stochastic contextual bandits with known reward functions. arXiv preprint arXiv:1605.00176.

Schramade, W. (2017). Investing in the UN Sustainable Development Goals: Opportunities for Companies and Investors, *Journal of Applied Corporate Finance*, 29: 87-99. DOI:10.1111/jacf.12236

Scott, Louise and Alan McGill. 2019. Creating a strategy for a better world. Pricewaterhouse Coopers. Accessed 2 July 2020: https://www.pwc.com/gx/en/sustainability/SDG/sdg-2019.pdf

Shestak, V.; Gura, D. and N. Khudyakova. (2020). Chatbot design issues: building intelligence with the Cartesian paradigm. *Evol. Intel.* DOI: https://doi.org/10.1007/s12065-020-00358-z

Singh, Om Prakash. (2019). Chatbots in psychiatry: Can treatment gap be lessened for psychiatric disorders in India. *Indian journal of psychiatry* vol. 61:3: 225. DOI:10.4103/0019-5545.258323

Sullivan, Hannah and Scott Schweikart. (2019). Are Current Tort Liability Doctrines Adequate for Addressing Injury Caused by AI?*AMA Journal of Ethics*, 21(2): E160-166. DOI: 10.1001/amajethics.2019.160.

Sustainalytics (2018) Managing data privacy risk: comparing the FAANG+ stocks. *ESG Spotlight*. June 7. Accessed 2 July 2020: https://www.sustainalytics.com/esg-research/issue-spotlights/faang-data-privacy/

Tatman, Rachael. (2016). Google's Speech Recognition Has a Gender Bias, *Making Noise and Hearing Things*, July 12. Accessed 2 July 2020: https://makingnoiseandhearingthings.com/2016/07/12/googles-speech-recognition-has-a-gender-bias/

Tesla Team. (2016). A Tragic Loss. *Tesla Blog*, June 30. Accessed 2 July 2020: https://www.tesla.com/blog/tragic-loss

Tett, Gillian. (2019). Ethical investing has reached a tipping point, London: *Financial Times*, June 18. Accessed 2 July 2020: https://www.ft.com/content/7d64d1d8-91a6-11e9-b7ea-60e35ef678d2

Thaler Robert and Cass Sunstein. (2008). *Nudge: Improving Decisions About Health, Wealth, and Happiness*. Yale University Press, New Haven.

Trucost. (2020). SDG Analytics for Investor Portfolios: Aligning Investments with Sustainability Goals. S&P Global. Accessed 2 July 2020: https://www.spglobal.com/_media/documents/sandp-trucost-sdg-impact-brochure.pdf

United Nations. (2015). United Nations Sustainable Development Goals. Accessed 2 July 2020: https://www.un.org/sustainabledevelopment/sustainable-development-goals/

(USBLS) United States Bureau of Labor Statistics. (2020). *Union Members Summary*. United States Department of Labor. Wednesday, January 22. Accessed 2 July 2020: https://www.bls.gov/news.release/union2.nr0.htm

Verma, Sahil and Julia Rubin. (2018). Fairness Definitions Explained. *Proceedings of the International Workshop on Software Fairness* (FairWare '18). Association for Computing Machinery, New York, NY, USA, 1–7. DOI: https://doi.org/10.1145/3194770.3194776

Vinuesa, R., Azizpour, H., Leite, I. Balaam, M., Dignum, V, Domisch, S., Felländer, A., Daniela, S., Tegmark, M., and Francesco Fuso Nerini. (2020) The role of artificial intelligence in achieving the Sustainable Development Goals. *Nature Communications* 11, 233. DOI: https://doi.org/10.1038/s41467-019-14108-y

Vinuesa, Ricardo; Theodorou, Andreas; Battaglini, Manuela and Virginia Dignum. (2020b). Comment: A socio-technical framework for digital contact tracing. Cornell University arXiv.org > cs > arXiv:2005.08370. Accessed 2 July 2020: https://arxiv.org/ftp/arxiv/papers/2005/2005.08370.pdf

Wang, Weiyu and Keng Siau. (2018). Ethical and Moral Issues with AI - A Case Study on Healthcare Robots. Twenty-fourth Americas Conference on Information Systems, New Orleans. Accessed 2 July 2020: https://www.researchgate.net/profile/Keng_Siau/publication/325934375_Ethical_and_Moral_Issues_with_AI/links/5b97316d92851c78c418f7e4/Ethical-and-Moral-Issues-with-AI.pdf

Washington, Anne. (2018). How to Argue with an Algorithm: Lessons from the COMPAS-ProPublica Debate, *Colorado Technology Law Journal*:17-1. Accessed 2 July 2020: https://ctlj.colorado.edu/?page_id=635

West, Sarah. (2019). Data Capitalism: Redefining the Logics of Surveillance and Privacy, *Business & Society*, 58(1), 20–41. DOI: https://doi.org/10.1177/0007650317718185

Westin, Alan. (1968). *Privacy and Freedom* (fifth ed.). New York: Atheneum.

Whittlestone, J. Nyrup, R. Alexandrova, A. Dihal, K. and S Cave. (2019). Ethical and societal implications of algorithms, data, and artificial intelligence: a roadmap for research. London: Nuffield Foundation.

Zerilli, J., Knott, A. and J Maclaurin. (2019). Transparency in Algorithmic and Human Decision-Making: Is There a Double Standard? *Philosophy and Technology* 32, 661–683. DOI: https://doi.org/10.1007/s13347-018-0330-6

Zuboff, Shoshana. (2019). *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. New York: Hatchette.