Philosophy & Technology (2022) 35:55
https://doi.org/10.1007/s13347-022-00553-z

RESEARCH ARTICLE

A Virtue-Based Framework to Support Putting AI Ethics into Practice

Thilo Hagendorff[1]

Received: 10 March 2022 / Accepted: 11 June 2022

Abstract
Many ethics initiatives have stipulated sets of principles and standards for good technology development in the AI sector. However, several AI ethics researchers have pointed out a lack of practical realization of these principles. Following that, AI ethics underwent a practical turn, but without deviating from the principled approach. This paper proposes a complementary to the principled approach that is based on virtue ethics. It defines four "basic AI virtues", namely justice, honesty, responsibility and care, all of which represent specific motivational settings that constitute the very precondition for ethical decision making in the AI field. Moreover, it defines two "second-order AI virtues", prudence and fortitude, that bolster achieving the basic virtues by helping with overcoming bounded ethicality or hidden psychological forces that can impair ethical decision making and that are hitherto disregarded in AI ethics. Lastly, the paper describes measures for successfully cultivating the mentioned virtues in organizations dealing with AI research and development.

Keywords AI virtues · AI ethics · Business ethics · Moral psychology · Bounded ethicality · Implementation · Machine learning · Artificial intelligence

1 Introduction

Current AI ethics initiatives, especially when adopted in scientific institutes or companies, mostly embrace a principle-based approach (Mittelstadt, 2019). However, establishing principles alone does not suffice; they also must be convincingly put into practice. Most AI ethics guidelines do shy away from coming up with methods to accomplish this (Hagendorff, 2020). Nevertheless, recently more and more research papers appeared that describe steps on how to come "from what to how" (Eitel-Porter, 2020; Morley et al., 2020; Theodorou & Dignum, 2020; Vakkuri et al.,

✉ Thilo Hagendorff
thilo.hagendorff@uni-tuebingen.de

[1] Cluster of Excellence "Machine Learning: New Perspectives for Science", University of Tuebingen, Tübingen, Germany

Published online: 21 June 2022

2019a). However, AI ethics still fails in certain regards. The reasons for that are manifold. This is why both in academia and public debates, many authors state that AI ethics has not permeated the AI industry yet, quite the contrary (Vakkuri et al., 2019b). Despite the mentioned reasons, this is due to current AI ethics discourses hardly taking considerations on moral psychology into account. They do not consider the limitations of the human mind, the many hidden psychological forces like powerful cognitive biases, blind spots and the like that can affect the likelihood of ethical or unethical behavior. In order to effectively improve moral decision making in the AI field and to live up to common ideals and expectations, AI ethics initiatives can seek inspiration from another ethical framework that is yet largely underrepresented in AI ethics, namely virtue ethics. Instead of focusing only on principles, AI ethics can put a stronger focus on virtues or, in other words, on character dispositions in AI practitioners in order to effectively put itself into practice. When using the term "AI practitioners" or "professionals", this includes AI or machine learning researchers, research project supervisors, data scientists, industry engineers and developers, as well as managers and other domain experts.

Moreover, to bridge the gap between existing AI ethics initiatives and the requirements for their successful implementation, one should consider insights from moral psychology because, up to now, most parts of the AI ethics discourse disregard the psychological processes that limit the goals and effectiveness of ethics programs. This paper aims to respond to this gap in research. AI ethics, in order to be truly successful, should not only repeat bullet points from the numerous ethics codes (Jobin et al., 2019). It should also discuss the right dispositions and character strengths in AI practitioners that can help not only to identify ethical issues and to engender the motivation to take action, but also—and this is even more important—to discover and circumvent one's own vulnerability to psychological forces affecting moral behavior. The purpose of this paper is to state how this can be executed and how AI ethics can choose a virtue-based approach in order to effectively put itself into practice.

2 AI Ethics—the Current Principled Approach

Current AI ethics programs often come with specific weaknesses and shortcomings. First and foremost, without being accompanied by binding legal norms, their normative principles lack reinforcement mechanisms (Rességuier & Rodrigues, 2020). Basically, deviations from codes of ethics have no or very minor consequences. Moreover, even when AI applications fulfill all ethical requirements stipulated, it does not necessarily mean that the application itself is "ethically approved" when used in the wrong contexts or when developed by organizations that follow unethical intentions (Hagendorff, 2021a; Lauer, 2020). In addition to that, ethics can be used for marketing purposes (Floridi, 2019; Wagner, 2018). Recent AI ethics initiatives of the private sector have faced a lot of criticism in this regard. In fact, industry efforts for ethical and fair AI are compared to past efforts of "Big Tobacco" to whitewash the image of smoking (Abdalla & Abdalla, 2020). "Big Tech", so the argument, uses ethics initiatives and targeted research funds to avoid legislation or the creation of binding legal norms (Ochigame, 2019). Hence, avoiding or addressing criticism like that is paramount for trustworthy ethics initiatives.

The latest progress in AI ethics research was configured by a "practical turn", which was among other things inspired by the conclusion that principles alone cannot guarantee ethical AI (Mittelstadt, 2019). To accomplish that, so the argument, principles must be put into practice. Recently, several frameworks were developed, describing the process "from what to how" (Hallensleben et al., 2020; Morley et al., 2020; Zicari, 2020). Basically, this implies considering the context dependency in the process of realizing codes of ethics, the different requirements for different stakeholders, as well as the demonstration of ways of dealing with conflicting principles or values, for instance in the case of fairness and accuracy (Whittlestone et al., 2019). Ultimately, however, the practical turn frameworks are often just more detailed codes of ethics that use more fine-grained concepts than the initial high-level guidelines. For instance, instead of just stressing the importance of privacy, like the first generation of comprehensive AI ethics guidelines did, they hint to the Privacy by Design or Privacy Impact Assessment toolkits (Cavoukian, 2011; Cavoukian et al., 2010; Oetzel & Spiekermann, 2014). Or instead of just stipulating principles for AI, they differentiate between stages of algorithmic development, namely business and use-case development; design phase, where the business or use case is translated into tangible requirements for AI practitioners; training and test data procurement; building of the AI application; testing the application; deployment of the application and monitoring of the application's performance (Morley et al., 2020). Other frameworks (Dignum, 2018) are rougher and differentiate between ethics by design (integrating ethical decision routines in AI systems (Hagendorff, 2021c)), ethics in design (finding development methods that support the evaluation of ethical implications of AI systems (Floridi et al., 2018)) and ethics for design (ensuring integrity on the side of developers (Johnson, 2017)). But, as stated above, all frameworks still stick to the principled approach. The main transformation lies in the principles being far more nuanced and less abstract compared to the beginnings of AI ethics code initiatives (Future of Life Institute, 2017). Typologies for every stage of the AI development pipeline are available. Differentiating principles solves one problem, namely the problem of too much abstraction. At the same time, however, it leaves some other problems open. Speaking more broadly, current AI ethics disregards certain dimensions it should actually be having. In organizations of all kinds, the likelihood of unethical decisions or behavior can be controlled to a certain extent. Antecedents for unethical behavior are individual characteristics (gender, cognitive moral development, idealism, job satisfaction, etc.), moral issue characteristics (the concentration and probability of negative effects, the magnitude of consequences, the proximity of the issue, etc.) and organizational environment characteristics (a benevolent ethical climate, ethical culture, code existence, rule enforcement, etc.) (Kish-Gephart et al., 2010). With regard to AI ethics, these factors are only partially considered. Most parts of the discourse are focused on discussing organizational environment characteristics (codes of ethics) or moral issues characteristics (AI safety) (Brundage et al., 2018; Hagendorff, 2020, 2021b), but not individual characteristics (character dispositions) increasing the likelihood of ethical decision making in AI research and development.

Therefore, a successful ethics strategy should focus on individual dispositions and organizational structures alike, whereas the overarching goal of every measure should be the prevention of harm. Or, in this case: prevent AI-based applications from inflicting direct or indirect harm. This rationale can be fulfilled by ensuring explainability of algorithmic decision making, by mitigating biases and promoting fairness in machine learning, by fostering AI robustness and the like. However, in addition to listing these issues is asking how AI practitioners can be taught to intuitively keep them in mind. This would mean to transition from a situation of an external "ethics assessment" of existing AI products with a "checkbox guideline" to an internal process of establishing "ethics for design".

Empirical research shows that having plain knowledge on ethical topics or moral dilemmas is likely to have no measurable influence on decision making. Even ethics professionals, meaning ethics professors and other scholars of ethics, typically do not act more ethically than non-ethicists (Schwitzgebel, 2009; Schwitzgebel & Rust, 2014). Correspondingly, in the AI field, empirical research shows that ethical principles have no significant influence on technology developer's decision making routines (McNamara et al., 2018). Ultimately, ethical principles do not suffice to secure prosocial ways to develop and use new technologies (Mittelstadt, 2019). Normative principles are not worth much if they are not acknowledged and adhered to. In order to actually acknowledge the importance of ethical considerations, certain character dispositions or virtues are required, among others, virtues that encourage us to stick to moral ideals and values.

3 Basic AI Virtues—the Foundation for Ethical Decision Making

Western virtue ethics has its roots in moral theories of Greek philosophers. However, after deontology and utilitarianism became more mainstream in modern philosophy, virtue ethics recently experienced a "comeback". Roughly speaking, this comeback of scholarly interest in virtue ethics was initiated by Anscombe's essay "Modern Moral Philosophy" (1958) but found prominent supporters and continued to grow by MacIntyre (1981), Nussbaum (1993), Hursthouse (2001) and many more. Virtue ethics also has a rich tradition in East and Southeast philosophy, especially in Confucian and Buddhist ethical theories (Keown, 1992; Tiwald, 2010). Virtue-based ethical theories treat character as fundamental to ethics, whereas deontology, arguably the most prevalent ethical theory, focusses on principles. But what are the differences between principles and virtues? The former is based on normative rules that are universally valid, the latter addresses the question of what constitutes a good person or character. While ethical principles equal obligations, virtues are ideals that AI practitioners can aspire to. Deontology-inspired normative principles focus on the action rather than the actor. Thus, principlism defines action-guiding principles, whereas virtue ethics demands the development of specific positive character dispositions or character strengths.

Why are these dispositions of importance for AI practitioners? One reason is that individuals, who display traits such as justice, honesty, empathy and the like, acquire (public) trust. Trust, in turn, makes it easier for people to cooperate and work together, it creates a sense of community and it makes social interactions more predictable (Schneier, 2012). Acquiring and maintaining the trust of other players in the AI field, but also the trust of the general public, can be a prerequisite for providing AI products and services. After all, intrinsically motivated actions are more trustworthy in comparison to those which are simply the product of extrinsically motivated rule following behavior (Meara et al., 1996).

One has to admit that a lot of ongoing AI basic research or very specific, small AI applications have such weak ethical implications that virtues or ethical values have no relevance at all. But AI applications that involve personal data, that are part of human–computer interaction or that are used on a grand scale clearly have ethical implications that can be addressed by virtue ethics. In the theoretical process of transitioning from an "uncultivated" to a morally habituated state, "technomoral virtues" like civility, courage, humility, magnanimity and others can be fostered and acquired (Vallor, 2016; Harris 2008a; Kohen et al., 2019; Gambelin, 2020; Sison et al., 2017; Neubert, 2017; Harris 2008b; Ratti & Stapleford, 2021). In philosophy, virtue ethics traditionally comprises cardinal virtues, namely fortitude, justice, prudence and moderation. Further, a list of six broad virtues that can be distilled from religious texts, oaths and other virtue inventories was put together by Peterson and Seligman (2004), whereas the virtues are wisdom, courage, humanity, justice, temperance and transcendence. Furthermore, in her famous book "Technology and the Virtues", Vallor (2016, 2021) identified twelve technomoral virtues, namely honesty, self-control, humility, justice, courage, empathy, care, civility, flexibility, perspective, magnanimity and wisdom. The selection was criticized in secondary literature (Howard, 2018; Vallor, 2018) but remains arguably the most important virtue-based approach in ethics of technology. In the more specific context of AI applications, however, one has to sort out those virtues that are particularly important in the field of AI ethics. Here, existing literature and preliminary works are spare (Constantinescu et al., 2021; Neubert & Montañez, 2020).

Based on patterns and regularities of the ongoing discussion on AI ethics, an ethics strategy that is based on virtues would constitute four basic AI virtues, where each virtue corresponds to a set of principles (see Table 1). The basic AI virtues are justice, honesty, responsibility and care. But how exactly can these virtues be derived from AI ethics principles? Why do exactly these four virtues suffice? When consulting meta-studies on AI ethics guidelines that stem from the sciences, industry, as well as governments (Fjeld et al., 2020; Hagendorff, 2020; Jobin et al., 2019), it becomes clear that AI ethics norms comprise a certain set of reoccurring principles. The mentioned meta-studies on AI ethics guidelines list these principles hierarchically, starting with the most frequently mentioned principles (fairness, transparency, accountability, etc.) and ending at principles that are mentioned rather seldom, but nevertheless repeatedly (sustainability, diversity, social cohesion etc.). When sifting through all these principles, one can, by using a reductionist approach and clustering them into groups, distill four basic virtues that cover all of them (see Fig. 1). The decisive question for the selection of the four basic AI virtues was: Does virtue *A* describe character dispositions that, when internalized by AI practitioners, will intrinsically motivate them to act in a way that "automatically" ensures or makes it more likely that the outcomes of their actions, among others, result in technological

Table 1 List of basic AI virtues

Basic AI virtues	Explanation	Corresponding principles
Justice	A strong sense of justice enables individuals to act fairly, meaning that they refrain from having any prejudice or favoritism towards individuals based on their intrinsic or acquired traits in the context of decision making. In AI ethics, justice is the one moral value that seems to be prioritized the most. However, it is hitherto operationalized mainly in mathematical terms, not with regard to actual character dispositions of AI practitioners. Here, justice as a virtue could not just underpin motivations to develop fair machine learning algorithms, but also efforts to use AI techniques only in those societal contexts where it is fair to apply them. Eventually, justice affects algorithmic non-discrimination and bias mitigation in data sets as well as efforts to avoid social exclusion, fostering equality and ensuring diversity	Algorithmic fairness, non-discrimination, bias mitigation, inclusion, equality, diversity
Honesty	Honesty is at the core of fulfilling a set of very important AI specific ethical issues. It fosters not only organizational transparency, meaning to provide information about financial or personnel related aspects regarding AI development. It also promotes the willingness to provide explainability or technical transparency regarding AI applications, for instance by disclosing origins of training data, quality checks the data were subject to, methods to find out how labels were defined etc. Moreover, honesty enables to acknowledge errors and mistakes that were made in AI research and development, allowing for collective learning processes	Organizational transparency, openness, explainability, interpretability, technological disclosure, open source, acknowledge errors and mistakes

Table 1 (continued)

Basic AI virtues	Explanation	Corresponding principles
Responsibility	For the AI sector, responsibility is of great importance and stands in place of a host of other positive character dispositions. Mainly, responsibility builds the precondition for feeling accountable for AI technologies and their outcomes. This is particularly relevant since AI technology's inherent complexity leads to responsibility diffusions that exacerbate the assignment of wrongdoing. Diffusions of responsibility in complex technological as well as social networks can cause individuals to detach themselves from moral obligations, possibly leading to breeding grounds for unethical behavior. Responsibility, seen as a character disposition, is a counterweight to that since it leads professionals to actually feeling liable for what they are doing, opposing negative effects of a diffusion of responsibility	Responsibility, liability, accountability, replicability, legality, accuracy, considering (long-term) technological consequences
Care	Care means to develop a sense for others' needs and the will to address them. Care has a strong connection to empathy, which is the precondition for taking the perspective of others and understanding their feelings and experiences. This way, care and empathy facilitate prosocial behavior and, on the other hand, discourage individuals from doing harm. In AI ethics, care builds the bedrock for motivating professionals to avoid AI applications from causing direct or indirect harm, ensuring safety, security, but also privacy preserving techniques. Moreover, care can motivate AI practitioners to design AI applications in a way that they foster sustainability, solidarity, social cohesion, common good, peace, freedom and the like. Care can be seen as being the driving force of the beneficial AI movement	Non-maleficence, harm, security, safety, privacy, protection, precaution, hidden costs, beneficence, well-being, sustainability, peace, common good, solidarity, social cohesion, freedom, autonomy, liberty, consent

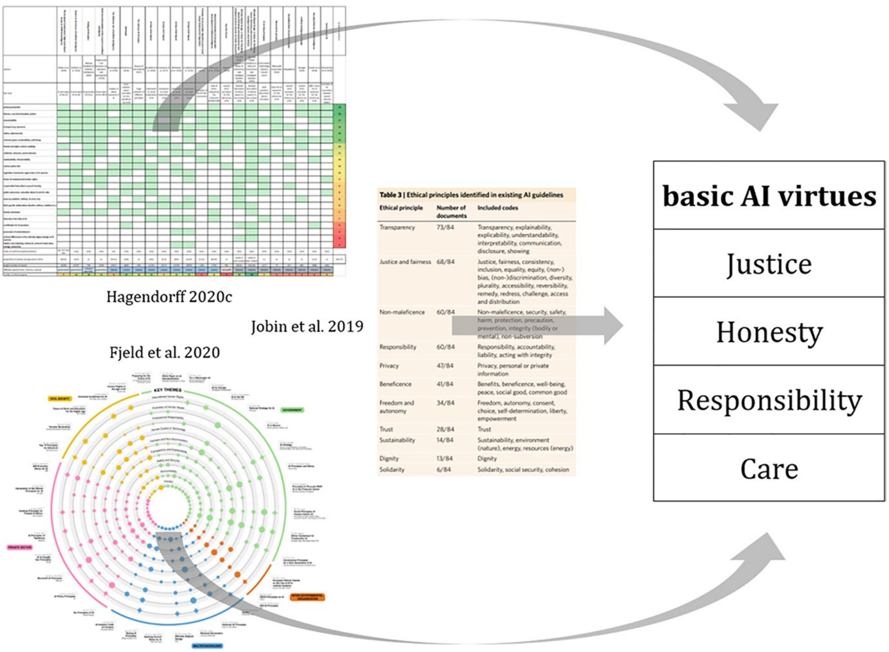

Table 3 | Ethical principles identified in existing AI guidelines

Ethical principle	Number of documents	Included codes
Transparency	73/84	Transparency, explainability, explicability, understandability, interpretability, communication, disclosure, showing
Justice and fairness	68/84	Justice, fairness, consistency, inclusion, equality, equity, (non-)bias, (non-)discrimination, diversity, plurality, accessibility, reversibility, remedy, redress, challenge, access and distribution
Non-maleficence	60/84	Non-maleficence, security, safety, harm, protection, precaution, prevention, integrity (bodily or mental), non-subversion
Responsibility	60/84	Responsibility, accountability, liability, acting with integrity
Privacy	47/84	Privacy, personal or private information
Beneficence	41/84	Benefits, beneficence, well-being, peace, social good, common good
Freedom and autonomy	34/84	Freedom, autonomy, consent, choice, self-determination, liberty, empowerment
Trust	28/84	Trust
Sustainability	14/84	Sustainability, environment (nature), energy, resources (energy)
Dignity	13/84	Dignity
Solidarity	6/84	Solidarity, social security, cohesion

Fig. 1 Using meta-studies on AI ethics guidelines as sources to distill four basic AI virtues

artefacts that meet the requirements that principle *X* specifies? Or, in short, does virtue *A* translate into behavior that is likely to result in an outcome that corresponds to the requirements of principle *X*? This question had to be applied for every principle that was derived from the meta-studies, testing by how many different virtues they can be covered. Ultimately, this process resulted in only four distinct virtues.

To name some examples: The principle of algorithmic fairness corresponds to the virtue of justice. A just person will "automatically" be motivated to contribute to machine outputs that do not discriminate against groups of people, independently of external factors and guideline rules. The principle of transparency, as a second example, corresponds to the virtue of honesty, because an honest person will "automatically" be inclined to be open about mistakes, to not hide technical shortcomings, to make research outcomes accessible and explainable. The principle of safe AI would be a third example. Here, the virtue of care will move professionals to act in a manner that they do not only acknowledge the importance of safety and harm avoidance, but also act accordingly. Ultimately, the transition happens between deontological rules, principles or universal norms on the one hand and virtues, intrinsic motives or character dispositions on the other hand. Nevertheless, both fields are connected by the same objective, namely to come up with trustworthy, human-centered, beneficial AI applications. Just the means to reach this objective are different.

As said before, the four basic AI virtues cover all common principles of AI ethics as described in prior discourses (Fjeld et al., 2020; Floridi et al., 2018; Hagendorff, 2020; Jobin et al., 2019; Morley et al., 2020). They are the precondition for putting principles into practice by representing different motivational settings for steering decision making processes in AI research and development in the right direction. But stipulating those four basic AI virtues is not enough. Tackling ethics problems in practice also needs second-order virtues that enable professionals to deal with "bounded ethicality".

4 Second-Order AI Virtues—a Response to Bounded Ethicality

When using a simple ethical theory, one can assume that individuals go through three phases. First, individuals perceive that they are confronted with a moral decision they have to make. Secondly, they reflect on ethical principles and come up with a moral judgment. And finally, they act accordingly to these judgments and therefore act morally. But individuals do not actually behave this way. In fact, moral judgments are in most cases not influenced by moral reasoning (Haidt, 2001). Moral judgments are done intuitively, and moral reasoning is used in hindsight to justify one's initial reaction. In short, typically, moral action precedes moral judgment. This leads to consequences for AI ethics. It shows that parts of current ethics initiatives can be reduced to plain "justifications" for the status quo of technology development—or at least they are adopted to it. For instance, the most commonly stressed AI ethics principles are fairness, accountability, explainability, transparency, privacy and safety (Hagendorff, 2020). However, these are issues for which a lot of technical solutions already exist and where a lot of research is done anyhow. Hence, AI ethics initiatives are simply reaffirming existing practices. On a macro level, this stands in correspondence with the aforementioned fact that moral judgments do not determine, but rather follow or explain prior decision making processes.

Although explicit ethics training may improve AI practitioners' intellectual understanding of ethics itself, there are many limitations restricting ethical decision making in practice, no matter how comprehensive one's knowledge on ethical theories is. Many reasons for unethical behavior are resulting from environmental influences on human behavior and limitations through bounded rationality or, to be more precise, "bounded ethicality" (Bazerman & Tenbrunsel, 2011; Tenbrunsel & Messick, 2004). Bounded ethicality is an umbrella term that is used in moral psychology to name environmental as well as intrapersonal factors that can thwart ethical decision making in practice. Hence, in order to address bounded ethicality, AI ethics programs are in need of specific virtues, namely virtues that help to "debias" ethical decision making in order to overcome bounded ethicality.

The first step to successively dissolve bounded ethicality is to inform AI practitioners not about the importance of machine biases, but psychological biases as well as situational forces. Here, two second-order virtues come into play, namely prudence and fortitude (see Table 2). In Aristotelian virtue ethics, prudence (or phrónēsis) guides the enactment of individual virtues in unique moral situations, meaning that a person can intelligently express virtuous behavior (Aristotle et al., 2012). As a unifying intellectual virtue, prudence also gains center stage in modern virtue-based approaches to engineering ethics (Frigo et al., 2021). In this paper, prudence plays a similar role and is used in combination with another virtue, namely

Table 2 List of second-order AI virtues

AI virtues	Explanation	Bounded ethicality
Prudence	Prudence means practical wisdom. In some philosophical theories, it represents the ability to gauge and reconcile complex and often competing values and requirements. Here, it stands for a high degree of self-understanding, for the ability to identify effects of bounded ethicality on one's own behavior as well as for the sincerity to acknowledge one's own vulnerability to unconscious cognitive biases. Prudence is the counterweight to the common limitations of the human mind, to the hidden psychological forces that impair ethical reasoning and decision making	System 1 thinking, implicit biases, in-group favoritism, self-serving biases, value-action gaps, moral disengagement, etc
Fortitude	Fortitude means idealism or the will to stick to moral ideals and moral responsibilities, potentially against all odds. For the AI sector, this means that researchers and managers acquire the courage to speak up when they come across moral issues. This may sound obvious, but in light of powerful situational forces, peer influences or authorities, speaking up and truly acting in accordance to one's own convictions can become very difficult. Fortitude helps to overcome these difficulties	Situational forces, peer influences, authorities, etc

fortitude. While both virtues may help to overcome bounded ethicality, they are at the same time enablers for living up to the basic virtues. Individual psychological biases as well as situational forces can get in the way of acting justly, honestly, responsibly or caringly. Prudence and fortitude are the answers to the many forces that may restrict basic AI virtues, where prudence is aiming primarily at individual factors, while fortitude addresses supra-individual issues that can impair ethical decision making in AI research and development.

In the following, a selection of some of the major factors of bounded ethicality that can be tackled by prudence shall be described. This selection is neither exhaustive nor does it go into much detail. However, it is meant to be a practical overview that can set the scene for more in-depth subsequent analyses.

Clearly, the most obvious factors of bounded ethicality are psychological biases (Cain & Detsky, 2008). It is common that people's first and often only reaction to moral problems is emotional. Or, in other words, taking up dual-process theory, their reaction follows *system 1 thinking* (Kahneman, 2012; Tversky & Kahneman, 1974), meaning an intuitive, implicit, effortless, automatic mode of mental information processing. System 1 thinking predominates everyday decisions. System 2, on the other hand, is a conscious, logical, less error-prone, but slow and effortful mode of thinking. Although many decision making routines would require system 2 thinking, individuals often lack the energy to switch from system 1 to system 2. Ethical decision making needs cognitive energy (Mead et al., 2009). This is why prudence is such an important virtue, since it helps AI practitioners to transition from system 1 to system 2 thinking in ethical problems. This is not to say that the dual-process theory is without criticism. Recently, cognitive scientists have challenged its validity (Grayot, 2020), even though they did not abandon it in toto. It still remains a scientifically sound heuristic in moral psychology. Thus, system 2 thinking remains strikingly close to critical ethical thinking, although it does obviously not necessarily result in it (Bonnefon, 2018).

The transition from system 1 to system 2 thinking in ethical problems can also be useful for mitigating another powerful psychological force, namely *implicit biases* (Banaji & Greenwald, 2013), that can impair at least two basic AI virtues, namely justice and care. Individuals have implicit associations, also called "ordinary prejudices", that lead them to classify, categorize and perceive their social surroundings with accordance to prejudices and stereotypes. This effect is so strong that even individuals who are absolutely sure to not be hostile towards minority groups actually are exactly that. The reason for that lies in the fact that people succumb to subconscious biases that reflect culturally established stereotypes or discrimination patterns. Hence, unintentional discrimination cannot be unlearned without changing culture, the media, the extent of exposure to people from minorities and the like. Evidently, this task cannot be fulfilled by the AI sector. Nevertheless, implicit biases can be tackled by increasing workforce diversity in AI firms and by using prudence as a virtue to accept the irrefutable existence and problematic nature of implicit biases as well as their influence on justice in the first place.

Another important bias that can compromise basic AI virtues and that can at the same time be overcome by prudence is *in-group favoritism* (Efferson et al., 2008). This bias causes people to sympathize with others who share their culture, organization, gender, skin color, etc. For AI practitioners, this means that AI applications which have negative side-effects on outgroups, for instance the livelihoods of clickworkers in South-east Asia (Graham et al., 2017), are rated less ethically problematic than AI applications that would have similar consequences for in-groups. Moreover, the current gender imbalance in the AI field might be prolonged by in-group favoritism in human resource management. In-group favoritism mainly stifles character dispositions like justice and care. Prudence, on the other hand, is apt to work against in-group favoritism by recognizing artificial group constructions as well as definitions of who counts as "we" and who as "others", bolstering not only fair decision making, but also abilities to empathize with "distant" individuals.

One further and important effect of bounded ethicality that can impair the realization of the basic AI virtues is *self-serving biases*. These biases cause revisionist impulses in humans, helping to downplay or deny past unethical actions while memorizing ethical ones, resulting in a self-concept that depicts oneself as ethical.

When one asks individuals to rate how ethical they think they are on a scale of 0 to 100 related to other individuals, the majority of them will give themselves a score of more than 50 (Epley & Dunning, 2000). The same holds true when people are asked to assess the organization they are a part of in relation to other organizations. Average scores are higher than 50, although actually the average score would have to be 50. What one can learn from this is that generally speaking, people overestimate their ethicality. Moreover, self-serving biases cause people to blame other people when things go wrong, but to view successes as being one's own achievement. Others are to blame for ethical problems, depicting the problems as being outside of one's own control. In the AI sector, self-serving biases can come into play when attributing errors or inaccuracies in applications as being the result of others, when reacting dismissive to critical feedback or feelings of concern, etc. Moreover, not overcoming self-serving biases by prudence can mean to act unjustly and dishonestly, further compromising basic AI virtues.

Value-action gaps are another effect of bounded ethicality revealed by empirical studies in moral psychology (Godin et al., 2005; Jansen & Glinow, 1985). Value-action gaps occur in the discrepancy between people's self-concepts or moral values and their actual behavior. In short, the gaps mark the distance between what people say and what people do. Prudence, on the other hand, can help to identify that distance. In the AI field, value-action gaps can occur on an organizational level, for instance by using lots of ethics-related terms in corporate reports and press releases while actually being involved in unethical businesses practices, lawsuits, fraud, etc. (Loughran et al., 2009). Especially the AI sector is often accused of ethics-washing, hence of talking much about ethics, but not acting accordingly (Hao, 2019). Likewise, value-action gaps can occur on an individual level, for instance by holding AI safety or data security issues in high esteem while actually accepting improper quality assurance or rushed development and therefore provoking technical vulnerabilities in machine learning models. Akin to value-action gaps are behavioral forecasting errors (Diekmann et al., 2003). Here, people tend to believe that they will act ethically in a given situation *X*, while when situation *X* actually occurs, they do not behave accordingly (Woodzicka & LaFrance, 2001). They underestimate the extent to which they will indeed stick to their ideals and intentions. All these effects can interfere negatively with basic AI virtues, mostly with care, honesty and justice. This is why prudence with regard to value-action gaps is of great importance.

The concept of *moral disengagement* is another important factor in bounded ethical decision making (Bandura, 1999). Techniques of moral disengagement allow individuals to selectively turn their moral concerns on and off. In many day-to-day decisions, people act contrary to their own ethical standards, but without feeling bad about it or having a guilty conscience. The main techniques in moral disengagement processes comprise justifications, where wrongdoing is justified as means to a higher end; changes in one's definition about what is ethical; euphemistic labels, where individuals detach themselves from problematic action contexts by using linguistic distancing mechanisms; denial of being personally responsible for particular outcomes, where responsibility is attributed to a larger group of people; the use of comparisons, where own wrongdoings are relativized by pointing at other contexts of wrongdoings or the avoidance of certain information that refers to negative consequences of one's own behavior. Again, prudence can help to identify cases of moral disengagement in the AI field and act as a response to it. Addressing moral disengagement with prudence can be a requirement to live up to all basic AI virtues.

In the following, a selection of some of the major factors of bounded ethicality that can be tackled by fortitude shall be described. Here, supra-individual issues that can impair ethical decision making in AI research and development are addressed. Certainly, one of the most relevant factors one has to discuss in this context are *situational forces*. Numerous empirical studies in moral psychology have shown that situational forces can have a massive impact on moral behavior (Isen & Levin, 1972; Latané & Darley, 1968; Williams & Bargh, 2008). Situational forces can range from specific influences like the noise of a lawnmower that significantly affects helping behavior (Mathews & Canon, 1975) to more relevant factors like competitive orientations, time constraints, tiredness, stress, etc., which are likely to alter or overwrite ethical concerns (Cave & ÓhÉigeartaigh, 2018; Darley & Batson, 1973; Kouchaki & Smith, 2014). Especially financial incentives have a significant influence on ethical behavior. In environments that are structured by economic imperatives, decisions that clearly have an ethical dimension can be reframed as pure business decisions. All in all, money has manifold detrimental consequences for decision making since it leads to decisions that are proven to be less social, less ethical or less cooperative (Gino & Mogilner, 2014; Gino & Pierce, 2009; Kouchaki et al., 2013; Palazzo et al., 2012; Vohs et al., 2006). Ultimately, various finance law obligations or monetary factual constraints that a company's management has to comply to can conflict with or overwrite AI virtues. Especially in contexts like this, virtue ethics can significantly be pushed into the background, although the perceived constraints lead to immoral outcomes. In short, situational forces can have negative impacts on unfolding all four basic AI virtues, namely justice, honesty, responsibility and care. In general, critics of virtue ethics have pointed out that moral behavior is not determined by character traits, but social contexts and concrete situations (Kupperman, 2001). However, situationist accounts are in fact entirely compatible with virtue ethics since it provides particular virtues like fortitude that are intended to counteract situational forces (and that can explain why some individuals deviate from expected behavior in classical psychological experiments like the Milgram experiment (Milgram, 1963)). Fortitude is supposed to help to counteract situational pressure, allowing the mentioned basic virtues to flourish.

Similar to and often not clearly distinguishable from situational forces are *peer influences* (Asch, 1951, 1956). Individuals want to follow the crowd, adapt their behavior to that of their peers and act similarly to them. This is also called conformity bias. Conformity biases can become a problem for two reasons: First, group norms can possess unethical traits, leading for instance to a collective acceptance of harm. Second, the reliance on group norms and the associated effects of conformity bias induces a suppression of own ethical judgments. In other words, if one individual starts to misbehave, for instance by cheating, others follow suit (Gino et al., 2009). A similar problem occurs with *authorities* (Milgram, 1963). Humans have an internal tendency for being obedient to authorities. This willingness to please authorities can have positive consequences when executives act ethically themselves. If this is not the case, the opposite becomes true. For AI ethics, this means that social norms that tacitly emerge from AI practitioner's behavioral routines as well as managerial decisions can both bolster ethical as well as unethical working cultures. In the case of the latter, the decisive factor is the way individuals respond to inner normative conflicts with their surroundings. Do they act in conformity and obedience even if it means to violate basic AI virtues? Or do they stick to their dispositions and deviate from detrimental social norms or orders? Fortitude, one of the two second-order virtues, can ensure the appropriate mental strength to stick to the right intentions and behavior, be it in cases where everyone disobeys a certain law but oneself does not want to join in, where managerial orders instruct to bring a risky product to the market as fast as possible but oneself insists on piloting it before release or where under extreme time pressure one insists on devoting time to understand and analyze training data sets.

5 Ethics Training—AI Virtues Come into Being

In traditional virtue ethics concepts, virtues emerge from habitual, repeated and gradually refined practice of right and prudent actions (Aristotle et al., 2012). At first, specific virtues are encouraged and practiced by performing acts that are inspired by "noble" human role-models and that resemble other patterns, narratives or social models of the virtue in question. Later, virtues are refined by taking the particularity of given situations into account. Regarding AI virtues, the proceeding is not much different (Bezuidenhout & Ratti, 2021). However, cultivating basic and second-order AI virtues means achieving virtuous practice embedded in a specific organizational and cultural context. A virtuous practice requires some sort of moral self-cultivation that encompasses the acquirement of motivations or the will to take action, knowledge on ethical issues, skills to identify them and moral reasoning to make the right moral decisions (Johnson, 2017). One could reckon that especially aforementioned skills or motivations are either innate or the result of childhood education. But ethical dispositions can be changed by education in all stages of life, for instance by powerful experiences, virtuous leaders or a certain work atmosphere in organizations. To put it in a nutshell, virtues can be trained and taught in order to foster ethical decision making and to overcome bounded ethicality. Most importantly, if ethics training imparts only explicit knowledge (or ethical principles), this will very likely have no effect on behavior. Ethics training must also impart tacit knowledge, meaning skills of social perception and emotion that cause individuals to automatically feel and want the right thing in a given situation (Haidt, 2006, p. 160).

The simplest form of ethics programs comprise ethics training sessions combined with incentive schemes for members of a given organization that reward the abidance of ethical principles and punish their violation. These ethics programs have numerous disadvantages. First, individuals that are part of them are likely to only seek to perform well on behavior covered by exactly these programs. Areas that are not covered are neglected. That way, ethics programs can even increase unethical behavior by actually well-intended sanctioning systems (Gneezy & Rustichini, 2000). For instance, in case a fine is put on a specific unethical behavior, individuals who benefit from this behavior might simply weigh the advantage of the unethical behavior against the disadvantage of the fine. If the former outweighs the latter, the unethical behavior might even increase if a sanctioning system is in place. Ethical decisions would simply be reframed as monetary decisions. In addition to that, individuals can become inclined to trick incentive schemes and reward systems. Moreover, those programs solely focus on extrinsic motivators and do not change intrinsic dispositions and moral attitudes. All in all, ethics programs that comprise simple reward and sanctioning systems—as well as corresponding surveillance and monitoring mechanisms—are very likely to fail.

A further risk of ethics programs or ethics training are reactance phenomena. Reactance occurs when individuals protest against constraints of their personal freedoms. As soon as ethical principles restrict the freedom of AI practitioners doing their work, they might react to this restriction by trying to reclaim that very freedom by all means (Dillard & Shen, 2005; Dowd et al., 1991; Hong, 1992). People want to escape restrictions, thus the moment when such restrictions are put in place—no matter whether they are justified from an ethical perspective or not—people might start striving to break free from them. Ultimately, "forcing" ethics programs on members of an organization is not a good idea. Ethics programs should not be decoupled from the inner mechanisms and routines of an organization. Hence, in order to avoid reactance and to fit ethics programs into actual structures and routines of an organization, it makes sense to carefully craft specific, unique compliance measures that take particular decision processes of AI practitioners and managers into account. In addition to that, ethics programs can be implemented in organizations with delay. This has the effect of a "future lock-in" (Rogers & Bazerman, 2008), meaning that policies achieve more support, since the time delay allows for an elimination of the immediate costs of implementation, for individuals to prepare for the respective measures and for a recognition of their advantages.

Considering all of that, what measures can actually support AI practitioners and AI companies' managers to strengthen AI virtues? Here, again, insights from moral psychology as well as behavioral ethics research can be used (Hines et al., 1987; Kollmuss & Agyeman, 2002; Treviño et al., 2006, 2014) to catalogue measures that bolster ethical decision making as well as virtue acquisition (see Tables 3 and 4). The measures can be vaguely divided into those that tend to affect single individuals and those that bring about or relate to structural changes in organizations. The following Table 3 lists measures that relate to AI professionals on an individual level.

The following Table 4 lists systemic measures that affect organizations mainly on a structural level.

6 Discussion

Virtue ethics does not come without shortcomings. In general, it is criticized for focusing on the "being" rather than the "doing", meaning that virtue ethics is agent- and not act-centered. Moreover, critics fault that on the one hand, virtuous persons can perform wrong actions, and on the other hand, right actions can be performed by persons who are not virtuous. However, this is a truism that could easily be

Table 3 Individual measures that bolster ethical decision making and virtue acquisition

Measures related to individuals	Explanation
Knowledge about AI virtues	AI professionals must be familiar with the six AI virtues and know about their importance and implications
Knowledge about action strategies	Professionals have to learn how they can mitigate ethically relevant problems, for instance in the fields of fairness, robustness, explainability, but also in terms of organizational diversity, clickwork outsourcing, sustainability goals, etc
Locus of control	Professionals should have the perception that they themselves are able to influence and have a tangible impact on ethically relevant issues. This also supports a sense of responsibility, meaning that professionals hold themselves accountable for the consequences of their decision making
Public commitment	Professionals can explicitly communicate the willingness to take action in ethical challenges. Publicly committing to stick to particular virtues, ideals, intentions and moral resolutions causes individuals to feel strongly obliged to actually do so when encountering respective choices
Audits and discussion groups	With the help of colleagues, one can reflect and discuss professional choices, ethical issues or other concerns in one's daily routines in order to receive critical feedback. Furthermore, fictious ethical scenarios simulating particular contexts of decision making that professionals may face can be used. Apart from scenario trainings, organizations can grant professionals time for contemplation, allowing time to read texts, e.g. about moral psychology or ethical theory

transferred to other approaches in ethical theory, for instance by pointing at the fact that normative rules can be disregarded or violated by individuals or that individuals can perform morally right actions without considering normative rules. Another response to that critique stresses that it is one of virtue ethics' major strength to not universally define "right" and "wrong" actions. Virtue ethics can address the question of "eudaimonia" without fixating axiological concepts of what is "right". Further, virtue ethics is criticized by pointing at its missing "codifiability". Stipulating sets of virtues is arbitrary. However, this critique also holds true for every other ethical theory. Their very foundations are always arbitrary. All in all, many points of criticism that are brought into position in order to find faults in virtue ethics can equally be brought into position against other ethical theories, such as deontology or consequentialist ethics.

Moreover, a further point of critique concerns the lack of technical details of the AI virtues approach. AI practitioners can censure the fact that the approach seems to be even more disconnected from down-to-earth research and development than the former, principled AI ethics initiatives. They also lacked technical details in many places or, in cases they mentioned details, did so in a very shallow manner (Hagendorff, 2020). The AI virtues concept, however, contains zero references to technical details—but for a reason. It is naïve to believe that ethical research is apt for that at all. Apart from the fact that a lot of ethical issues cannot be solved by technical means in the first place, AI ethics is the wrong discipline to come up with technical

Table 4 Systemic measures that bolster ethical decision making and virtue acquisition

Systemic measures	Explanation
Leader influences	Managers play a key role as role models influencing employee's attitudes and behaviors. Their decisions have a particular legitimacy and credibility, which makes employees imitating them very likely. This way, managers, whose prosocial attitudes, fairness and behavioral integrity are of utmost importance, can define ethical standards in their organizations, since their way of making moral decisions trickles down to subordinate individuals (Treviño et al., 2014)
Ethical climate and culture	Unlike ethics codes, which are proven to have no significant effect on (un)ethical decision making in organizations, ethical climates do have that effect (McNamara et al., 2018). Especially caring climates are positively related to ethical behavior. On the other hand, self-interested, egoistic climates are negatively associated with ethical choice. Furthermore, ethical cultures, meaning informal norms, language, rituals etc., also affect ethical decision making and can, among other things, be significantly influenced by performance management systems (Kish-Gephart et al., 2010)
Proportion of women	Countless studies in empirical business ethics research indicate that women are more sensitive to and less tolerant of unethical activities than their male counterparts (Loe et al., 2013). In short, gender is a predictor to ethical behavior. This points out the importance of raising the proportion of female employees. Especially in the AI sector, male researchers currently strikingly outnumber females. This lack of workforce diversity has consequences on the functionality of software applications as well as implications on ethical outcomes in AI organizations. Hence, raising the proportion of women in the AI sector should pose one of the most effective measures to improve ethical decision making on a grand scale. This is not to say that the same does not hold true for other underrepresented demographics or marginalized populations. Here, the paper only points at the hiring of women, though. This is due to the fact that only for women and not for other demographic groups, ample research shows that they are less likely to engage in unethical behavior compared to men
Decreasing stress and pressure	Reducing the amount of stress and time pressure in organizations can have game-changing consequences for the organizations' ethical climates (Darley & Batson, 1973; Selart & Johansen, 2011). De-stressing professionals, slowing down processes and, by that, setting cognitive resources free promote a transition from system 1 to system 2 thinking in decision making situations. This way, simply speaking, individuals are encouraged to think before they act, which can ultimately improve ethicality in organizations
Openness for critique	Critical voices from the public can point at blind spots or specific shortcomings of organizations. Being open to embrace external critique as an opportunity to reflect upon an organization's own routines and goals with the associated willingness to potentially realign them can significantly improve its own trustworthiness, reputation and public perception. Eventually, this can contribute to the overall success of an organization

details on privacy preserving machine learning, explainability, sustainable model training, etc. Instead, AI practitioners themselves are the ones who can do that. But they also need to be motivated to consult literature, tools and frameworks on these technical details. And virtues are the basis for this motivation. The least thing ethics codes' principles can do is to point at particular technical papers or methods on how to achieve fair, safe, explainable, privacy preserving, etc. AI, but only virtues motivate to actually use these methods. Hence, it is not a weakness of the virtues framework presented in this paper to not contain any references to technical details—it is, in fact, the expression of an appropriate unpretentiousness about its competencies.

Another critical issue a virtue-based ethics framework must address is the fact that it focuses only on AI practitioners and not a wider socio-economic context or systemic changes. Regarding the latter, ethics discourses can play an important role in inspiring laws or political objectives. However, ethics as a philosophical enterprise that involves the study of principles, values or virtues cannot unfold the same efficacy as binding legal norms that comprise concrete duties, that are able to resolve disputes and that are established by democratic institutions. Hence, when talking about efficiency gains in applying ethics or about ethics' practical turn, this should never cause the impression that ethics acquires a similar or even the same steering effect or enforceability that binding legal norms or similar systemic measures possess. Ultimately, trustworthy AI will be the result of both strands, ethics as well as law. Both strands interact and inspire each other. However, especially virtue ethics with its focus on individual dispositions is perhaps less apt to inspire systemic changes or legal norms than principlism.

In addition to that, the problem of unethical AI usage is not per se caused by individuals in research and development. Cases in which AI applications cause harm can result from multifactorial, dynamic events that are not directly intended by anyone. Unforeseen technological consequences cannot be attributed to a lack of virtuous behavior. One could argue that one of the basic AI virtues, namely care, also implies the willingness to assess long-term technological consequences, but this does obviously not guarantee that harmful technological consequences can be strictly avoided. It would put too much responsibility on individual AI practitioners to blame them for all ethical issues that are tied to the use of AI. Within the general scheme of things, AI practitioners are powerful, but also not omnipotent players who are accompanied by many other agents who have direct or indirect influences on AI technologies. Responsibilities are in many cases widely shared between groups of AI practitioners, meaning researchers, engineers, managers and other domain experts. However, this distribution of responsibilities does not mean that they somehow vanish. It is a known effect of moral disengagement that responsibility diffusion can cause individuals to detach themselves from moral obligations. The virtue-based framework presented here is supposed to counteract this.

Another shortcoming one has to discuss in the context of a virtue-based approach in AI ethics revolves around effects of elitism. When AI practitioners, once educated in the basic and second-order AI virtues, become solely responsible for their actions, who will have moral authority over them? Or, when extending the scope of this question, one can also ask: What authority does the author of the framework presented in this paper have to say what virtues practitioners should develop? A possible reply would be to refer to discourse ethics where communicative rationality is used to agree on the validity of particular moral norms or, in this case, virtues (Habermas, 2001). However, this paper follows this methodology only in a very indirect, remote manner. It derives the authority for the selection of virtues from the fact that they are not the result of subjective preferences, but meta-studies on AI ethics that are by themselves the result of a global discourse on AI ethics. However, "global" in this case is somewhat misleading, since the geographic distribution of origin countries of AI ethics guidelines is rather biased towards economically developed countries (Jobin et al., 2019). Hence, especially African and South-American countries are not represented in the AI ethics discourse. This also has an influence on the selection of the presented AI virtues, meaning that they are likely to have a tendency to represent a Western perspective.

7 Conclusion

Hitherto, all the major AI ethics initiatives choose a principled approach. They aim at having an effect on AI research and development by stipulating a list of rules and standards. But, as more and more papers from AI-metaethics show (Hagendorff, 2020; Lauer, 2020; Mittelstadt, 2019; Rességuier & Rodrigues, 2020), this approach has specific shortcomings. The principled approach in AI ethics has no reinforcement mechanisms, it is not sensitive to different contexts and situations, it sometimes fails to address the technical complexity of AI, it uses terms and concepts that are often too abstract to be put into practice, etc. In order to improve the two last-named shortcomings, AI ethics recently underwent a practical turn, stressing its will to put principles into practice. But the typologies and guidelines on how to put AI ethics into practice stick to the principled approach altogether (Hallensleben et al., 2020; Morley et al., 2020). However, a hitherto largely underrepresented approach, namely virtue ethics, seems to be a promising addition to AI ethics' principlism.

The goal of this paper was to outline how virtues can support putting AI ethics into practice. Virtue ethics focuses on an individual's character development. Character dispositions provide the basis for professional decision making. On the one hand, the paper considered insights from moral psychology on the many pitfalls the motivation of moral behavior has. On the other hand, it used virtue instead of deontological ethics to promote and foster not only four basic AI virtues, but also two second-order AI virtues that can help to circumvent "bounded ethicality" and one's vulnerability to unconscious biases. The basic AI virtues comprise justice, honesty, responsibility and care. Each of these virtues motivates a kind of professional decision making that builds the bedrock for fulfilling all the AI specific ethics principles discussed in literature. In addition to that, the second-order AI virtues, namely prudence and fortitude, can be used to overcome specific effects of bounded ethicality that can stand in the way of the basic AI virtues, meaning biases, value-action gaps, moral disengagement, situational forces, peer influences and the like. Lastly, the paper described framework conditions and organizational measurements that can help to realize ethical decision making and virtue training in the AI field. Equipped with this information, organizations dealing with AI research and development should be able to effectively put AI ethics into practice.

Author contribution TH is the sole author.

Funding Open Access funding enabled and organized by Projekt DEAL. This research was supported by the Cluster of Excellence “Machine Learning—New Perspectives for Science” funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany’s Excellence Strategy—Reference Number EXC 2064/1—Project ID 390727645.

Data Availability Not applicable.

Declarations

Ethics Approval and Consent to Participate Not applicable

Consent for Publication Not applicable

Competing Interests The author declares no competing interests.

References

Abdalla, M., & Abdalla, M. (2020). The Grey Hoodie Project: Big Tobacco, Big Tech, and the threat on academic integrity. *arXiv*, 1–9.

Anscombe, G. E. M. (1958). Modern moral philosophy. *Philosophy, 33*(124), 1–19.

Aristotle, Barlett, R. C., & Collins, S. D. (2012). *Aristotle’s Nicomachean ethics*. University of Chicago Press.

Asch, S. (1951). Effects of group pressure upon the modification and distortion of judgment. In H. S. Guetzkow (Ed.), *Groups, leadership and men*: *Research in human relations* (pp. 177–190). Pittsburgh: Russell & Russell.

Asch, SE. (1956). Studies of independence and conformity: I. A minority of one against a unanimous majority. *Psychological Monographs: General and Applied, 70*(9), 1–70.

Banaji, M. R., & Greenwald, A. G. (2013). *Blindspot: Hidden biases of good people*. Delacorte Press.

Bandura, A. (1999). Moral disengagement in the perpetration of inhumanities. *Personality and Social Psychology Review, 3*(3), 193–209.

Bazerman, M. H., & Tenbrunsel, A. E. (2011). *Blind spots: Why we fail to do what’s right and what to do about it*. Princeton University Press.

Bezuidenhout, L., & Ratti, E. (2021). What does it mean to embed ethics in data science? An integrative approach based on microethics and virtues. *AI & SOCIETY - Journal of Knowledge, Culture and Communication, 36*(3), 939–953.

Bonnefon, J.-F. (2018). The pros and cons of identifying critical thinking with system 2 processing. *Topoi, 37*(1), 113–119.

Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., et al. (2018). The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. *arXiv*, 1–101.

Cain, D. M., & Detsky, A. S. (2008). Everyone's a little bit biased (even physicians). *JAMA, 299*(24), 2893–2895.

Cave, S., & ÓhÉigeartaigh, S. S. (2018). An AI race for strategic advantage: Rhetoric and risks 1–5.

Cavoukian, A. (2011). Privacy by design: The 7 foundational principles: Implementation and mapping of fair information practices. https://iapp.org/media/pdf/resource_center/Privacy%20by%20Design%20-%207%20Foundational%20Principles.pdf. Accessed 21 June 2018.

Cavoukian, A., Taylor, S., & Abrams, M. E. (2010). Privacy by design: Essential for organizational accountability and strong business practices. *Identity in the Information Society, 3*(2), 405–413.

Constantinescu, M., Voinea, C., Uszkai, R., & Vică, C. (2021). Understanding responsibility in responsible AI. Dianoetic virtues and the hard problem of context. *Ethics and Information Technology, 23*(4), 803–814.

Darley, J. M., & Batson, C. D. (1973). "From Jerusalem to Jericho": A study of situational and dispositional variables in helping behavior. *Journal of Personality and Social Psychology, 27*(1), 100–108.

Diekmann, K. A., Tenbrunsel, A. E., & Galinsky, A. D. (2003). From self-prediction to self-defeat: Behavioral forecasting, self-fulfilling prophecies, and the effect of competitive expectations. *Journal of Personality and Social Psychology, 85*(4), 672–683.

Dignum, V. (2018). Ethics in artificial intelligence: Introduction to the special issue. *Ethics and Information Technology, 20*(1), 1–3.

Dillard, J. P., & Shen, L. (2005). On the nature of reactance and its role in persuasive health communication. *Communication Monographs, 72*(2), 144–168.

Dowd, E. T., Milne, C. R., & Wise, S. L. (1991). The therapeutic reactance scale: A measure of psychological reactance. *Journal of Counseling & Development, 69*(6), 541–545.

Efferson, C., Lalive, R., & Fehr, E. (2008). The coevolution of cultural groups and ingroup favoritism. *Science, 321*(5897), 1844–1849.

Eitel-Porter, R. (2020). Beyond the promise: Implementing ethical AI. *AI and Ethics*, 1–8.

Epley, N., & Dunning, D. (2000). Feeling "holier than thou": Are self-serving assessments produced by errors in self- or social prediction? *Journal of Personality and Social Psychology, 79*(6), 861–875.

Fjeld, J., Achten, N., Hilligoss, H., Nagy, A., & Srikumar, M. (2020). Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for AI. Berkman Klein Center Research Publication No. 2020–1. *SSRN Electronic Journal*, 1–39.

Floridi, L. (2019). Translating principles into practices of digital ethics: Five risks of being unethical. *Philosophy & Technology, 32*(2), 185–193.

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., et al. (2018). AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds and Machines, 28*(4), 689–707.

Frigo, G., Marthaler, F., Albers, A., Ott, S., & Hillerbrand, R. (2021). Training responsible engineers. Phronesis and the role of virtues in teaching engineering ethics. *Australasian Journal of Engineering Education, 26*(1), 25–37.

Future of Life Institute. (2017). Asilomar AI principles. Future of life institute. https://futureoflife.org/ai-principles/. Accessed 23 October 2018.

Gambelin, O. (2020). Brave: What it means to be an AI ethicist. *AI and Ethics*, 1–5.

Gino, F., Ayal, S., & Ariely, D. (2009). Contagion and differentiation in unethical behavior: The effect of one bad apple on the barrel. *Psychological Science, 20*(3), 393–398.

Gino, F., & Mogilner, C. (2014). Time, money, and morality. *Psychological Science, 25*(2), 414–421.

Gino, F., & Pierce, L. (2009). The abundance effect: Unethical behavior in the presence of wealth. *Organizational Behavior and Human Decision Processes, 109*(2), 142–155.

Gneezy, U., & Rustichini, A. (2000). A fine is a price. *The Journal of Legal Studies, 29*(1), 1–17.

Godin, G., Conner, M., & Sheeran, P. (2005). Bridging the intention-behaviour 'gap': The role of moral norm. *The British Journal of Social Psychology, 44*(Pt 4), 497–512.

Graham, M., Hjorth, I., & Lehdonvirta, V. (2017). Digital labour and development: Impacts of global digital labour platforms and the gig economy on worker livelihoods. *Transfer: European Review of Labour and Research, 23*(2), 135–162.

Grayot, J. D. (2020). Dual process theories in behavioral economics and neuroeconomics: A critical review. *Review of Philosophy and Psychology, 11*(1), 105–136.

Habermas, J. (2001). *Moral consciousness and communicative action*. Cambridge (Mass.): MIT.

Hagendorff, T. (2020). The ethics of AI ethics: An evaluation of guidelines. *Minds and Machines, 30*(3), 457–461.

Hagendorff, T. (2021a). Blind spots in AI ethics. *AI and Ethics*, 1–17.

Hagendorff, T. (2021). Forbidden knowledge in machine learning: Reflections on the limits of research and publication. *AI & SOCIETY - Journal of Knowledge, Culture and Communication, 36*(3), 767–781.

Hagendorff, T. (2021). Linking human and machine behavior: A new approach to evaluate training data quality for beneficial machine learning. *Minds and Machines, 31*, 563–593.

Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychology Review, 108*(4), 814–834.

Haidt, J. (2006). *The happiness hypothesis: Putting ancient wisdom and philosophy to the test of modern science*. Arrow Books.

Hallensleben, S., Hustedt, C., Fetic, L., Fleischer, T., Grünke, P., Hagendorff, T., et al. (2020). From principles to practice: An interdisciplinary framework to operationalise AI ethics. Gütersloh: Bertelsmann Stiftung, 1–56.

Hao, K. (2019). In 2020, let's stop AI ethics-washing and actually do something. https://www.technologyreview.com/s/614992/ai-ethics-washing-time-to-act/. Accessed 7 January 2020.

Harris, C. E. (2008). The good engineer: Giving virtue its due in engineering ethics. *Science and Engineering Ethics, 14*(2), 153–164.

Hines, J. M., Hungerford, H. R., & Tomera, A. N. (1987). Analysis and synthesis of research on responsible environmental behavior: A meta-analysis. *The Journal of Environmental Education, 18*(2), 1–8.

Hong, S.-M. (1992). Hong's psychological reactance scale: A further factor analytic validation. *Psychological Reports, 70*(2), 512–514.

Howard, D. (2018). Technomoral civic virtues: A critical appreciation of Shannon Vallor's technology and the virtues. *Philosophy & Technology, 31*(2), 293–304.

Hursthouse, R. (2001). *On virtue ethics*. Oxford University Press.

Isen, A. M., & Levin, P. F. (1972). Effect of feeling good on helping: Cookies and kindness. *Journal of Personality and Social Psychology, 21*(3), 384–388.

Jansen, E., & von Glinow, M. A. (1985). Ethical ambivalence and organizational reward systems. *The Academy of Management Review, 10*(4), 814–822.

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence, 1*(9), 389–399.

Johnson, D. G. (2017). Can engineering ethics be taught? *The Bridge, 47*(1), 59–64.

Kahneman, D. (2012). *Thinking, fast and slow*. Penguin.

Keown, D. (1992). *The nature of Buddhist ethics*. Palgrave MacMillan.

Kish-Gephart, J. J., Harrison, D. A., & Treviño, L. K. (2010). Bad apples, bad cases, and bad barrels: Meta-analytic evidence about sources of unethical decisions at work. *The Journal of Applied Psychology, 95*(1), 1–31.

Kohen, A., Langdon, M., & Riches, B. R. (2019). The making of a hero: Cultivating empathy, altruism, and heroic imagination. *Journal of Humanistic Psychology, 59*(4), 617–633.

Kollmuss, A., & Agyeman, J. (2002). Mind the gap: Why do people act environmentally and what are the barriers to pro-environmental behavior? *Environmental Education Research, 8*(3), 239–260.

Kouchaki, M., & Smith, I. H. (2014). The morning morality effect: The influence of time of day on unethical behavior. *Psychological Science, 25*(1), 95–102.

Kouchaki, M., Smith-Crowe, K., Brief, A. P., & Sousa, C. (2013). Seeing green: Mere exposure to money triggers a business decision frame and unethical outcomes. *Organizational Behavior and Human Decision Processes, 121*(1), 53–61.

Kupperman, J. J. (2001). The indispensability of character. *Philosophy, 76*(296), 239–250.

Latané, B., & Darley, J. M. (1968). Group inhibition of bystander intervention in emergences. *Journal of Personality and Social Psychology, 10*(3), 215–221.

Lauer, D. (2020). You cannot have AI ethics without ethics. *AI and Ethics*, 1–5.

Loe, T. W., Ferrell, L., & Mansfield, P. (2013). A review of empirical studies assessing ethical decision making in business. In A. C. Michalos & D. C. Poff (Eds.), *Citation classics from the journal of business ethics* (pp. 279–301). Springer, Netherlands.

Loughran, T., McDonald, B., & Yun, H. (2009). A wolf in sheep's clothing: The use of ethics-related terms in 10-K reports. *Journal of Business Ethics, 89*(S1), 39–49.

MacIntyre, A. C. (1981). *After virtue: A study in moral theory*. University of Notre Dame Press.

Mathews, K. E., & Canon, L. K. (1975). Environmental noise level as a determinant of helping behavior. *Journal of Personality and Social Psychology, 32*(4), 571–577.

McNamara, A., Smith, J., & Murphy-Hill, E. (2018). Does ACM's code of ethics change ethical decision making in software development? In G. T. Leavens, A. Garcia, & C. S. Păsăreanu (Eds.) (pp. 1–7). New York,: ACM Press.

Mead, N. L., Baumeister, R. F., Gino, F., Schweitzer, M. E., & Ariely, D. (2009). Too tired to tell the truth: Self-control resource depletion and dishonesty. *Journal of Experimental Social Psychology, 45*(3), 594–597.

Meara, N. M., Schmidt, L. D., & Day, J. D. (1996). Principles and virtues. *The Counseling Psychologist, 24*(1), 4–77.

Milgram, S. (1963). Behavioral study of obedience. *Journal of Abnormal Psychology, 67*, 371–378.

Mittelstadt, B. (2019). Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence, 1*(11), 501–507.

Morley, J., Floridi, L., Kinsey, L., & Elhalal, A. (2020). From what to how. An overview of AI ethics tools, methods and research to translate principles into practices. *Science and Engineering Ethics, 26*, 2141–2168.

Neubert, M. J. (2017). Teaching and training virtues: Behavioral measurement and pedagogical approaches. In A. J. G. Sison, G. R. Beabout, & I. Ferrero (Eds.), *Handbook of virtue ethics in business and management* (pp. 647–655). Springer, Netherlands.

Neubert, M. J., & Montañez, G. D. (2020). Virtue as a framework for the design and use of artificial intelligence. *Business Horizons, 63*(2), 195–204.

Nussbaum, M. (1993). Non-relative virtues: An Aristotelian approach. In M. Nussbaum & A. Sen (Eds.), *The quality of life* (pp. 242–269). Oxford University Press.

Ochigame, R. (2019). The invention of "ethical AI": How big tech manipulates academia to avoid regulation. https://theintercept.com/2019/12/20/mit-ethical-ai-artificial-intelligence/. Accessed 7 January 2020.

Oetzel, M. C., & Spiekermann, S. (2014). A systematic methodology for privacy impact assessments: A design science approach. *European Journal of Information Systems, 23*(2), 126–150.

Palazzo, G., Krings, F., & Hoffrage, U. (2012). Ethical blindness. *Journal of Business Ethics, 109*(3), 323–338.

Peterson, C., & Seligman, M. E. P. (2004). *Character strengths and virtues: A handbook and classification*. American Psychological Association.

Ratti, E., & Stapleford, T. A. (Eds.). (2021). *Science, technology, and virtues: Contemporary perspectives*. Oxford University Press.

Rességuier, A., & Rodrigues, R. (2020). AI ethics should not remain toothless! A call to bring back the teeth of ethics. *Big Data & Society, 7*(2), 1–5.

Rogers, T., & Bazerman, M. H. (2008). Future lock-in: Future implementation increases selection of 'should' choices. *Organizational Behavior and Human Decision Processes, 106*, 1–20. https://doi.org/10.1016/j.obhdp.2007.08.001

Schneier, B. (2012). *Liars & outliers: Enabling the trust that society needs to thrive*. John Wiley & Sons.

Schwitzgebel, E. (2009). Do ethicists steal more books? *Philosophical Psychology, 22*(6), 711–725.

Schwitzgebel, E., & Rust, J. (2014). The moral behavior of ethics professors: Relationships among self-reported behavior, expressed normative attitude, and directly observed behavior. *Philosophical Psychology, 27*(3), 293–327.

Selart, M., & Johansen, S. T. (2011). Ethical decision making in organizations: The role of leadership stress. *Journal of Business Ethics, 99*(2), 129–143.

Sison, AJG., Beabout, GR., & Ferrero, I. (Eds.). (2017). *Handbook of virtue ethics in business and management*. Dordrecht: Springer Netherlands.

Tenbrunsel, A. E., & Messick, D. M. (2004). Ethical fading: The role of self-deception in unethical behavior. *Social Justice Research, 17*(2), 223–236.

Theodorou, A., & Dignum, V. (2020). Towards ethical and socio-legal governance in AI. *Nature Machine Intelligence, 2*(1), 10–12.

Tiwald, J. (2010). Confucianism and virtue ethics: Still a fledgling in Chinese and comparative philosophy. *Comparative Philosophy: An International Journal of Constructive Engagement of Distinct Approaches toward World Philosophy, 1*, 2.

Treviño, L. K., den Nieuwenboer, N. A., & Kish-Gephart, J. J. (2014). (Un)ethical behavior in organizations. *Annual Review of Psychology, 65*, 635–660.

Treviño, L. K., Weaver, G. R., & Reynolds, S. J. (2006). Behavioral ethics in organizations: A review. *Journal of Management, 32*(6), 951–990.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*(4157), 1124–1131.

Vakkuri, V., Kemell, K-K., & Abrahamsson, P. (2019a). AI ethics in industry: A research framework. *arXiv*, 1–10.

Vakkuri, V., Kemell, K-K., Kultanen, J., Siponen, M., & Abrahamsson, P. (2019b). Ethically aligned design of autonomous systems: Industry viewpoint and an empirical study. *arXiv*, 1–17.

Vallor, S. (2016). *Technology and the virtues: A philosophical guide to a future worth wanting*. Oxford University Press.

Vallor, S. (2018). Technology and the virtues: A response to my critics. *Philosophy & Technology, 31*(2), 305–316.

Vallor, S. (2021). Twenty-first-century virtue: Living well with emerging technologies. In E. Ratti & T. A. Stapleford (Eds.), *Science, technology, and virtues: Contemporary perspectives* (pp. 77–96). Oxford University Press.

Vohs, K. D., Mead, N. L., & Goode, M. R. (2006). The psychological consequences of money. *Science, 314*(5802), 1154–1156.

Wagner, B. (2018). Ethics as an escape from regulation: From ethics-washing to ethics-shopping? In M. Hildebrandt (Ed.), *Being profiled: Cogitas ergo sum* (pp. 84–89). Amsterdam University Press.

Whittlestone, J., Nyrup, R., Alexandrova, A., & Cave, S. (2019). The role and limits of principles in AI ethics. In V. Conitzer, G. Hadfield, & S. Vallor (Eds.) (pp. 195–200). New York, NY, USA: ACM.

Williams, L. E., & Bargh, J. A. (2008). Experiencing physical warmth promotes interpersonal warmth. *Science, 322*(5901), 606–607.

Woodzicka, J. A., & LaFrance, M. (2001). Real versus imagined gender harassment. *Journal of Social Issues, 57*(1), 15–30.

Zicari, RV. (2020). Z-inspection: A holistic and analytic process to assess ethical AI. Mindful use of AI. http://z-inspection.org/wp-content/uploads/2020/10/Zicari.Lecture.October15.2020.pdf. Accessed 24 November 2020.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.