T

Anastasia Siapka

A Virtue Ethics Approach to AI-induced Risk

> Thinking clearly about risks and their acceptability in our lives is too important to be left to technical risk assessors and cost-benefit theorists[1].
>
> Carl Cranor

1. *From a technological risk society to risk-based technology regulation*

Once an uncontrollable force, tied to pre-determined natural or spiritual factors, risk has become a major societal preoccupation since the 20th century[2]. New technologies introduce risks that transcend spatial, temporal and social boundaries, ushering in a «risk society», where diverse yet incomprehensible futures are possible[3]. More recently, Artificial Intelligence (AI) systems – understood as (sets of) algorithms performing goal-oriented tasks that would otherwise require human intelligence – incur risks that necessitate attention and intervention[4]. Thus, self-regulatory frameworks addressed to AI compa-

[1] C.F. Cranor, *Toward a Non-Consequentialist Approach to Acceptable Risks*, in T. Lewens (ed.), *Risk: Philosophical Perspectives*, Routledge, London 2007, p. 51.

[2] J. van der Heijden, *Risk as an Approach to Regulatory Governance: An Evidence Synthesis and Research Agenda*, in «SAGE Open» 11, no. 3 (September 2021), pp. 1-12, https://doi.org/10.1177/21582440211032202.

[3] U. Beck, *Risk Society: Towards a New Modernity*, trans. M. Ritter, *Theory, Culture & Society*, Sage Publications, London 1992, p. 9; A. Giddens, *Risk Society: The Context of British Politics*, in J. Franklin (ed.), *The Politics of Risk Society*, Polity Press, Cambridge 1998, pp. 23-34.

[4] A. Siapka, *The Ethical and Legal Challenges of Artificial Intelligence: The EU Response to Biased and Discriminatory AI*, SSRN Scholarly Paper, Social Science Research Network, New York, 11 December 2018, https://dx.doi.org/10.2139/ssrn.3408773.

 DOI: 10.4454/wj35mz94

nies and developers are conceived specifically for or adapted to AI risk[5].

This pervasiveness of risk, hitherto confined to private practices, affects legally binding regulation. Under the General Data Protection Regulation (GDPR), AI developers acting as data controllers consider «risks of varying likelihood» and perform impact assessments for high-risk processing, but are provided with minimal guidance on how to do so[6]. The Artificial Intelligence Act (AIA) moves further than the GDPR does, adopting a «proportionate risk-based approach» as a core feature of its architecture[7]. AI developers acting as providers adhere to different obligations (e.g., conformity assessments, monitoring, risk management systems and voluntary codes of conduct) based on the system's risk level[8]. Despite the AIA's expansive material and territorial scope, including its possible role as a «benchmark» for other jurisdictions given the «Brussels effect», guidance on the risk-based approach remains vague, leaving AI developers «to their own devices»[9].

[5] Examples of the former include the NIST AI Risk Management Framework and ISO/IEC 23894, while an example of the latter is COSO ERM 201715: J. Schuett, *Risk Management in the Artificial Intelligence Act*, in «European Journal of Risk Regulation» 15, no. 2 (2024), pp. 368-369, https://doi.org/10.1017/err.2023.1.

[6] Articles 24 (1), 25 (1), 35 (1) and Recital 75. *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/EC (General Data Protection Regulation)*, Pub. L. No. 32016R0679, OJ L 119 (2016), http://data.europa.eu/eli/reg/2016/679/oj/eng.

[7] *Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts*, Pub. L. No. COM/2021/206 final (2021), https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:52021PC0206; Directorate General for Communication, *EU AI Act: First Regulation on Artificial Intelligence*, Article, European Parliament, Strasbourg (France), 19 December 2023, https://www.europarl.europa.eu/pdfs/news/expert/2023/6/story/20230601STO93804/20230601STO93804_en.pdf. In December 2023, the EU's Parliament and Council reached a provisional, political agreement on the contents of the long-awaited AIA. However, as at the time of writing the revised text has not been released, I take into account its 2021 version. Given that the paper does not go into detail about specific provisions and instead uses the AIA for illustrative purposes only, any subsequent changes to the text of the law are not expected to affect my arguments therein.

[8] A precursor to the AIA's approach is found in the work of the German Data Ethics Commission. In 2019, this Commission put forward a «risk-adapted regulatory approach», suggesting a classification of AI systems into five levels of criticality. Datenethikkommission, *Opinion of the Data Ethics Commission*, Data Ethics Commission of the Federal Government, Berlin, December 2019, pp. 173-182, https://www.bmi.bund.de/SharedDocs/downloads/EN/themen/it-digital-policy/datenethikkommission-abschlussgutachten-lang.pdf;jsessionid=789B1C3D1FC30ACF12B067AD01FDFD38.live881?__blob=publicationFile&v=5.

[9] Schuett, *op. cit.*, pp. 367-380. The «Brussels effect» describes to the EU's power to influence rules and regulations in other jurisdictions beyond its Member States. As for the AIA's

This rising prominence yet concurrent under-specification of risk-based approaches to AI constitutes the first reason for this paper's focus on AI risk[10].

The second reason concerns the multiple material and immaterial forms that AI risk takes[11]. Examples of the former include adverse outcomes to health and safety by AI-embedded products; impeded access to essential services by AI-based credit scoring; and deprivation of liberty by AI used in law enforcement[12]. Examples of the latter include discrimination by AI-based social scoring; surveillance and hindered freedom of assembly by AI used for remote biometric identification; and diminished career and education prospects when AI determines access to educational or employment opportunities[13]. Therefore, AI risk spans individuals, groups and society as a whole. These risks are posed by Narrow AI, which outperforms humans in specific tasks yet lacks the versatility of human intelligence[14]. Contrariwise, General AI (or Artificial General Intelligence) would be endowed with broad cognitive abilities tantamount to those of humans[15]. The mere possibility of General AI, especially after developments in Generative AI, has sparked concerns about longer-term, existential risks to humankind by systems unaligned with human values[16]. This paper does not further examine AI risks, but targets the approaches used for their assessment. Rejecting technocratic approaches, it evaluates AI risk through two contrasting normative theories: consequentialism and virtue ethics.

scope, it covers AI systems across multiple application domains and encompasses all providers placing AI on the market or putting it into service in the EU.

[10] This focus does not exclude the applicability of the paper's arguments to other types of technology. In addition, regulation is here understood in a broad sense, comprising regulatory acts by actors that may or may not have a legal mandate, in line with Black and Murray's definition (Section 4): «By regulation (and regulatory governance) is meant sustained and focused attempts to change the behaviour of others in order to address a collective problem or attain an identified end or ends, usually but not always through a combination of rules or norms and some means for their implementation and enforcement, which can be legal or non-legal». J. Black, A.D. Murray, *Regulating AI and Machine Learning: Setting the Regulatory Agenda*, in «European Journal of Law and Technology» 10, no. 3 (30 December 2019), https://www.ejlt.org/index.php/ejlt/article/view/722.

[11] Artificial Intelligence Act, *op. cit.*

[12] *Ibidem*.

[13] *Ibidem*.

[14] Siapka, *op. cit.*, pp. 17, 22.

[15] *Ibidem*.

[16] Generative AI implies AI systems that «generate brand-new, unique artifacts». Gartner, *Definition of Generative AI*, in «Gartner Glossary», Information Technology Glossary, accessed 25 August 2023, https://www.gartner.com/en/information-technology/glossary/generative-ai. For an overview of approaches to the existential risk (or x-risk) of AI, see, PauseAI, *The Existential Risk of Superintelligent AI*, in «Pause AI», accessed 16 December 2023, https://pauseai.info/xrisk.

2. *The technocratic approach to risk*

a. *Objectivity vs. normativity*

By and large, (self-)regulatory instruments divide risk-based approaches into two stages: (i) risk assessment, implying the identification of risks and evaluation of their acceptability and (ii) risk management, including the selection and adoption of measures to mitigate the previously identified and evaluated risks[17]. The first stage is considered an objective, neutral process, in which technical expert advice leaves little to no room for normative judgement[18]. The normative character of the second stage is more straightforward, since decisions about risk mitigation involve not only scientific and technical but also ethical, societal, political, financial, practical and other qualitative considerations.

However, this distinction between a value-free process of risk assessment and a normative one of risk management is artificial[19]. Far from being discovered by experts in an exclusively empirical way, risks are identified also on the basis of norms, values and often subjective perceptions, while being «strongly involved with social relations and meanings»[20]. As argued in science and technology studies and in the foundational report *Taking European Knowledge Society Seriously* specifically, «questions of risk can be recognised intrinsically to be shaped and framed by social values, sometimes embodied in routinised habitual ways of institutional thinking, and political interests»[21]. Indicatively, selecting the forms of risk relevant to the assessment, the measurement criteria to be employed, the weight to be placed on possible effects, and the thresholds of risk acceptability is a value-laden process[22]. Focusing on certain dimensions of risk privileges some normative

[17] A third stage of risk communication may also be distinguished but is not strictly relevant to the arguments of this paper.

[18] U. Felt *et al.*, *Taking European Knowledge Society Seriously*, Report of the Expert Group on Science and Governance to the Science, Economy and Society Directorate, Directorate General for Research, European Commission, Directorate General for Research and Innovation (European Commission), Belgium, January 2007, pp. 32-42, https://op.europa.eu/en/publication-detail/-/publication/5d0e77c7-2948-4ef5-aec7-bd18efe3c442; C.F. Cranor, *The Normative Nature of Risk Assessment: Features and Possibilities*, in «RISK: Health, Safety & Environment (1990-2002)» 8, no. 2 (March 1997), pp. 123-136; N. van Dijk, R. Gellert, K. Rommetveit, *A Risk to a Right? Beyond Data Protection Risk Assessments*, in «Computer Law & Security Review» 32, no. 2 (April 2016), pp. 286-306, https://doi.org/10.1016/j.clsr.2015.12.017; van der Heijden, *op. cit.*

[19] Felt *et al.*, *op. cit.*, pp. 32-42.

[20] van Dijk, Gellert, Rommetveit, *op. cit.*, p. 289.

[21] Felt *et al.*, *op. cit.*, p. 34.

[22] *Ibidem*.

perspectives or commitments while occluding others. Therefore, risk assessments are, at least in part, normatively construed.

This omission of normativity matters beyond risk assessment. It affects risk management, which consecutively builds upon and reflects the types of risk identified during assessment. It also affects the risk-based approach more broadly, given its function in facilitating decision-making about risks. Granted that risk-based approaches aim to eliminate or reduce risks, if these are not accurately identified and evaluated in the stage of assessment, given its disregard for normativity, the measures adopted for such elimination or reduction in the subsequent stage of management will be correspondingly misguided. This interconnection between risk assessment and management is so strong that the possibility of their separate treatment is doubted[23]. Hence, risk-based approaches, be they in voluntary or binding regulation, fail to achieve their aims unless they incorporate both objective and normative considerations throughout.

b. *Technical vs. ethical understanding*

Risk is broadly a «technique for creating knowledge and certainty about future events that are uncertain by definition»[24]. EU legal instruments associate it with the notions of «likelihood» or «probability» of harm and its «severity»[25]. These notions point to a technical understanding of risk, nu-

[23] Cranor, *Normative Nature of Risk Assessment*, *op. cit.* p. 128.

[24] van Dijk, Gellert, Rommetveit, *op. cit.*, p. 301.

[25] See, respectively, «[a] risk is a scenario describing an event and its consequences, estimated in terms of severity and likelihood» and «severity and likelihood of this risk should be assessed» in Article 29 Data Protection Working Party, *Guidelines on Data Protection Impact Assessment (DPIA) and Determining Whether Processing Is "Likely to Result in a High Risk" for the Purposes of Regulation 2016/679*, European Commission, Brussels, 4 April 2017, p. 6, http://ec.europa.eu/newsroom/article29/item-detail.cfm?item_id=611236; *Opinion 05/2014 on Anonymisation Techniques*, European Commission, Brussels, 10 April 2014, p. 7, https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf. Similar wording is used in Recitals 75-76 GDPR: General Data Protection Regulation, *op. cit*. Likewise in the AIA, «the AI systems pose a risk of harm to the health and safety, or a risk of adverse impact on fundamental rights, that is, in respect of its severity and probability of occurrence» and «taking into account both the severity of the possible harm and its probability of occurrence»: Artificial Intelligence Act, *op. cit*. As for other EU legislation, «"risk" means a function of the probability of an adverse health effect and the severity of that effect, consequential to a hazard» in *Consolidated Text: Regulation (EC) No 178/2002 of the European Parliament and of the Council of 28 January 2002 Laying down the General Principles and Requirements of Food Law, Establishing the European Food Safety Authority and Laying down Procedures in Matters of Food Safety*, Pub. L. No. OJ L 031 (2002), http://data.europa.eu/eli/reg/2002/178/2019-07-26. Like-

merically representing the outcome of the probability of a possible harm multiplied by the severity of said harm.

From this perspective, risk is distinguished from uncertainty. In decisions under risk, the probabilities of different adverse outcomes materialising are available and part of the calculation of risk[26]. In decisions under uncertainty, the different possible outcomes might or might not be available, but their probabilities are definitely not[27]. This distinction is, however, contested. The exact probability of a risk occurring is known solely in artificial cases (e.g., rolling a dice), compared to the more frequent real-life cases of uncertainty, where the probabilities of possible outcomes are unknown[28]. Relying on statistics and probabilities, this understanding of risk simplifies «the full range of uncertainties to the more comforting illusion of controllable, probabilistic but deterministic processes»[29].

Conversely, ethicists invoke risk in its ordinary usage, denoting the possibility that an adverse or undesirable outcome, such as harm, injury or loss, will occur[30]. This broader view of risk illuminates nuances that the focus on probability and severity overlooks. Given that ethics examines the attribution of praise and blame, it approaches risks differently depending on whether they are apt for such an attribution. Hence, risks that we face differ from risks that we take[31]. The first type includes risks whose occurrence we cannot control but whose management is to a certain degree under our control (e.g., risks caused by natural disasters). The second type includes risks to which exposure is chosen and over which there is a dimension of control

wise, «"risk" means the probable rate of occurrence of a hazard causing harm and the degree of severity of the harm» in *Directive 2009/48/EC of the European Parliament and of the Council of 18 June 2009 on the Safety of Toys*, OJ L 170 § (2009), https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX:32009L0048.

[26] M. Hayenhjelm, J. Wolff, *The Moral Problem of Risk Impositions: A Survey of the Literature*, in «European Journal of Philosophy» 20 (June 2012), p. E30, https://doi.org/10.1111/j.1468-0378.2011.00482.x.

[27] *Ibidem*.

[28] S.O. Hansson, *Philosophical Perspectives on Risk*, in «Techné: Research in Philosophy and Technology» 8, no. 1 (Fall 2004), pp. 11-12, https://doi.org/10.5840/techne2004818.

[29] B. Wynne, *Uncertainty and Environmental Learning: Reconceiving Science and Policy in the Preventive Paradigm*, in «Global Environmental Change» 2, no. 2 (June 1992), p. 123, https://doi.org/10.1016/0959-3780(92)90017-2.

[30] K. Shrader-Frechette, *Risk*, in *Routledge Encyclopedia of Philosophy*, 1st ed., Routledge, London 1998, https://doi.org/10.4324/9780415249126-L088-1. On the criticism against technocratic approaches to risk, see van der Heijden, *op. cit*.

[31] N. Rescher, *Risk: A Philosophical Introduction to the Theory of Risk Evaluation and Management*, University Press of America, Washington 1983, pp. 6-7. Although in practice there might be overlapping or borderline cases, the distinction adds nuance to the moral picture.

lacking from the first type (e.g., risks caused by human-made products). Risks of the first type approximate incidents of luck, impeding ascriptions of responsibility. It is the second type, risks to which we decide to expose ourselves and others, that matters for responsibility. Within this second type, we can differentiate between risks to which we decide to expose ourselves (self-imposed) and those imposed on us by others (other-imposed)[32]. In the latter case, the roles of those imposing the risk, their motivations for doing so, and the voluntary or not acceptance of these externally imposed risks matter from an ethical standpoint yet are captured by neither severity nor probability.

Based on the foregoing, the technocratic approach to risk, comprising an objectivist perspective on risk assessment and a scientific conceptualisation of risk, is rejected as artificial and overly narrow. In this paper, AI risk is not a free-floating, objectively accessible and measurable entity whose assessment exclusively relies on the properties of the AI system in question. Instead, it is conceived in its ethical usage, denoting the possibility of a future undesirable event occurring because of AI development/deployment, and particularly in its second type, denoting risks that involve the exercise of choice by AI developers. The process of its assessment is likewise considered imbued with normativity.

3. *The consequentialist approach to risk*

a. *Overview*

Although ethicists acknowledge that a complete absence of risk is impossible, they seek to evaluate the extent to which risk is acceptable. In most cases, they do so by appealing to consequentialism[33]. Consequentialism is the strand of normative ethical theory that evaluates actions as morally right or wrong based on a comparison of their overall beneficial and harmful consequences.

Consequentialist approaches to risk are premised upon the assumption that all consequences are comparable and aggregable[34]. The standard form they take is the Risk Cost Benefit Analysis (RCBA)[35]. Regulatory agencies

[32] Cranor, *Toward a Non-Consequentialist Approach to Acceptable Risks*, *op. cit.*, p. 50.

[33] Hayenhjelm, Wolff, *op. cit.*, pp. E28, E32.

[34] S.O. Hansson, *Risk and Ethics: Three Approaches*, in T. Lewens (ed.), *Risk: Philosophical Perspectives*, Routledge, London 2007, p. 26.

[35] T. Lewens, *Introduction: Risk and Philosophy*, in T. Lewens (ed.), *Risk: Philosophical Perspectives*, Routledge, London 2007, pp. 1-20.

employ the RCBA to assess the desirability of varying technological interventions, «from building a liquefied natural gas facility to adding yellow dye number two to margarine»[36]. RCBA encompasses «decision-aiding techniques» that seek to identify all likely good (benefits) and bad (risks/costs) consequences of an option and, by employing numerical terms, to «add up the likely overall good consequences of a decision option and to subtract from that figure the likely overall bad consequences»[37]. If the resulting overall good/bad consequences ratio is favourable – i.e., if the former outweigh the latter – risk is acceptable. Upon repeating this process for all available options, the one maximising net benefits or minimising net risks/costs is chosen.

Comparisons between good and bad consequences are straightforward when these are of the same type – e.g., if AI decreases the jobs available in a certain domain but increases those available in another. This is not often the case, though, rendering such comparisons difficult. For example, AI deployment in healthcare may be concurrently linked to the benefit of faster access to treatment and to the risk of biased diagnoses. For this reason, such approaches convert consequences that may differ a lot from each other into a single, usually monetary attribute[38]. To achieve this conversion, RCBA identifies «how much people would be willing to pay to have (or to avoid) these consequences»[39]. Following the previous example, individuals' hypothetical willingness to pay more for faster AI-enabled medical treatment than for avoiding a racially biased AI-enabled diagnosis would suggest the acceptability of AI risk.

However, not all RCBA techniques are single-attribute ones. Multi-attribute risk benefit analysis suggests that, as a first step, each consequence should be measured separately using the scale appropriate for it[40]. In the previous example, the number of hours from admission to treatment might be appropriate for measuring the consequences of AI-enabled healthcare

[36] K. Shrader-Frechette, *The Real Risks of Risk-Cost-Benefit Analysis*, in «Technology in Society» 7, no. 4 (1985), p. 399, https://doi.org/10.1016/0160-791X(85)90007-7.

[37] Lewens, *op. cit.*, p. 7. See also S.O. Hansson, *Risk*, in E.N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy*, Metaphysics Research Lab, Stanford University, 2018, https://plato.stanford.edu/archives/fall2018/entries/risk/; Shrader-Frechette, *The Real Risks of Risk-Cost-Benefit Analysis*, *op. cit*.

[38] Hansson, *Risk*, *op. cit.*, section 7.4.

[39] Lewens, *op. cit.*, p. 5.

[40] M. Peterson, *On Multi-Attribute Risk Analysis*, in T. Lewens (ed.), *Risk: Philosophical Perspectives*, Routledge, London 2007, pp. 68-83. If we, however, consider the presence of a single scale to be a defining feature of risk/cost benefit analyses, then multi-attribute analysis can be deemed a distinct kind of consequentialist approach.

services, whereas the number of lives saved as a result of accurate AI-enabled diagnoses might be appropriate for measuring the risks of AI bias. As a next step, these measurements are aggregated to formulate an overall ranking for each action and then used to compare alternative actions based on their overall rankings[41]. Although multi-attribute analyses include more dimensions than single-attribute ones do, they resemble the latter in expressing consequences in aggregated, numerical terms.

b. *Objections*

Where legal instruments refer to weighing the risks of an option against its benefits, they allude to some sort of RCBA, as the seemingly rational and rigorous guidance of this approach has rendered it the dominant choice[42]. Despite their popularity and ostensible precision, however, consequentialist approaches to risk face objections.

First, accurately identifying the consequences brought about, for instance, by AI is possible only in hindsight, after these have come to fruition. Such approaches cannot assist developers in evaluating the system's outcomes beforehand. Alternatively, consequentialists appeal to expected or hypothetical, instead of actual, consequences. However, the novelty of emerging technologies, including AI, poses difficulties in predicting their consequences – even just the expected ones – and the probability of their occurrence. Comparing AI's risks and benefits demands considerable information; yet, it is questionable whether such information is at the developers' disposal for the time being and, even where that is the case, whether such information is sufficiently complete or reliable. Due to its unpredictable nature and fast pace of development, the consequences of AI are often unintentional and unexpected; asking developers to identify risks that are by their nature hard to foresee is admittedly a tall order.

Second, by maximising benefits over costs in the aggregate, consequentialist approaches are impersonal in terms of their distributive effects. They would favour AI systems that incur more benefits than costs, without examining who would bear these. In this way, they leave open the problematic possibility that risks/costs are piled up in one part of the population and

[41] *Ibidem*.

[42] For example, the proposed AIA prohibits remote biometric identification, except for narrowly defined cases in which the public interest benefits outweigh the risks. Artificial Intelligence Act, *op. cit*.

benefits in another[43]. Even if the risks shouldered by the first population group were extremely severe, they would be justified by consequentialism as long as the second population group, which would bear the benefits, were larger[44]. AI developers would thus evaluate risks and benefits to users generally, without distinguishing among the needs and characteristics of different individuals or groups as such or in comparison with each other. Nonetheless, not all individuals or groups experience risk in the same way. Certain groups (e.g., children) are considered more vulnerable and require particular attention, which consequentialism could not justify. Apart from differences across individuals/groups at a given time, differences might exist across generations (e.g., present ones embracing the benefits while future ones bear the risks), for which consequentialist calculations cannot account. The use of willingness-to-pay indicators likewise overlooks that these depend on one's income: those with lower incomes are inevitably able and thereby willing to pay less to avoid risk than those with higher incomes without this implying that the former are actually less risk averse[45].

Third, consequentialist approaches to risk are not merely impersonal but even crude or cruel. Examining solely the outcomes of actions, they overlook the means employed to reach said outcomes. If an AI system promised benefits that significantly overrode its costs, its development/deployment would be acceptable even if, for example, relevant decision-making occurred through authoritarian procedures. Relatedly, consequentialist approaches do not differentiate between risks we face and those we take nor do they account for risks imposed, the agent(s) imposing these risks, their motivations and the (in)voluntary acceptance by risk bearers, all aspects that section 2 considered morally significant.

The cruelty of consequentialism additionally emerges in its effort to homogenise all values, rights, goods and moral commitments by translating them into commensurable terms[46]. For instance, AI risks to human life

[43] This would be reminiscent of Beck's claim that «wealth accumulates at the top, risk at the bottom»: Beck, *op. cit*., p. 35.

[44] We might even conceive of a three-party relationship, in which one group is subject to risks/costs owing to transactions between two other benefitting groups.

[45] Shrader-Frechette, *The Real Risks of Risk-Cost-Benefit Analysis*, cit., p. 403. For a more detailed version of Shrader-Frechette's assessment of the RCBA, see K.S. Shrader-Frechette, *Assessing Risk-Cost-Benefit Analysis, The Preeminent Method of Technology Assessment and Environmental-Impact Analysis*, in *Science Policy, Ethics, and Economic Methodology: Some Problems of Technology Assessment and Environmental-Impact Analysis*, D. Reidel, Dordrecht 1985, pp. 32-64, https://doi.org/10.1007/978-94-009-6449-5_2.

[46] Shrader-Frechette, *The Real Risks of Risk-Cost-Benefit Analysis*, cit.

or the environment would be placed on the same (monetary) scale as AI's potential benefits in efficiency and would even be acceptable if that scale tilted towards the latter. There is something corrupting about the mere act of subjecting goods, such as human life or the environment, to such calculations. This crude approach to values or goods that are commonly considered «priceless or sacred» alters their perceived worth in harmful ways, converting them into tradeable commodities[47]. Put simply, «some goods are cheapened when we try to attach a price to them»[48].

A fourth and broader objection is metaphysical. It challenges the adequacy of consequentialist approaches in capturing the breadth of «human situational understanding»[49]. By focusing on «allegedly transparent rationality and scientific know-how», analytic, formal and economic frameworks of thought upon which RCBA draws are reductionist and detached from reality[50]. In practice, human decision-making (especially in policy) resists such formalisation. It relies on intuitions and judgements that, akin to wisdom, are shaped by expertise and skills beyond algorithmic ways of thinking[51]. Even attempts to engage in such formal ways of thinking are unlikely to succeed, as humans' perspectives on what might be the consequences of an action differ substantially, as do their perspectives on which of these consequences are good or bad. For instance, if AI deployment in logistics is likely to increase the number of product deliveries achieved within a certain timeframe, this likelihood might be classified as a cost/risk by an environmentalist but as a benefit by an economist.

4. *The case for an alternative approach*

a. *Overview*

Traditional ethical theories are geared towards evaluating actions with more or less certain or knowable outcomes (*deterministic bias*)[52]. When extended to non-determinate settings, meaning to actions whose outcomes are

[47] D. MacLean, *Cost-Benefit Analysis and Procedural Values*, in «Analyse & Kritik» 16, no. 2 (1994), p. 171, https://doi.org/10.1515/auk-1994-0205.

[48] MacLean, *op. cit.*, p. 168.

[49] Shrader-Frechette, *The Real Risks of Risk-Cost-Benefit Analysis*, cit., p. 400.

[50] *Ibidem*.

[51] *Ibidem*.

[52] S. O. Hansson, *Ethical Criteria of Risk Acceptance*, in «Erkenntnis» 59, no. 3 (2003), p. 291, https://doi.org/10.1023/A:1026005915919.

uncertain, or to mixed determinate and non-determinate settings, the result is unsatisfactory, as the preceding objections to consequentialism demonstrate.

If, however, as Cranor cautions in the epigraph and following my argument thus far, risk acceptability cannot be entrusted to technical or cost-benefit risk assessors, to whom should it be assigned? This paper suggests an examination of risk acceptability from a normative perspective that does not focus on the certain or uncertain outcomes of actions and might thereby evade the deterministic bias of mainstream ethics. I refer here to (Aristotelian) virtue ethics. Redirecting ethical enquiry from the question of «*what should I do?*» to «*what sort of person should I be?*», virtue ethics concentrates on one's character and specifically on whether it manifests virtues. Virtue is a stable disposition of a person to do the right thing for the right reasons, in the right way and with the right emotion. The right thing to do is a mean state between two possible reactions, an excessive and a deficient one, and differs according to the situation at hand.

A virtue-ethical approach, then, shifts the focus from the actual or expected consequences of the risk-inducing situation to the «the risk-taker as an intentional agent and, in particular, on said *agent's attitude towards risk-taking* and *sensitivity to the context* in which risks are taken, all of which will reflect her moral character»[53]. Although AI developers may not control the outcome of risk-inducing decisions, they do control the decisions to take risks, so they should be deemed responsible for these decisions. Following the distinction in section 2 between risks we face and those we take, given that AI is deliberately developed and deployed by humans, its risks approximate those to which we decide to expose ourselves and others, compared to, say, risks incurred by natural disasters. Hence, focusing moral evaluation and responsibility attribution on AI developers' character and their decision to risk, rather than on the consequences resulting from such a decision, seems justified. From this perspective, the fact that the adverse consequences of a risk-inducing AI system did not materialise (e.g., because of luck or other external factors) would not suffice to retrospectively absolve AI developers from their responsibility if their decision-making was vicious. Conversely, that their actions eventually led to the imposition of risk or harm would not suffice to affirm their responsibility if their overall attitude was virtuous.

[53] N. Athanassoulis, A. Ross, *A Virtue Ethical Account of Making Decisions about Risk*, in «Journal of Risk Research» 13, no. 2 (March 2010), p. 218, https://doi.org/10.1080/13669870903126309. Emphasis added.

b. *Relation to consequentialism*

Because of this shift towards developers' character and decision-making, virtue ethics evades the first objection to consequentialism about AI's uncertain consequences. Being preoccupied with the «*what should I do?*» question, consequentialism is exclusively act-centred. By contrast, as a predominantly (yet not exclusively) agent-centred theory, virtue ethics embraces an open-ended reflexivity, which takes into account the situational particulars of normative problems and thereby of emerging technologies. By engaging the agent's reasoning, virtue ethics is better placed to address diverse and borderline ethical issues that are inadequately subsumed under binary comparisons or inflexible calculations.

Concerning the objections about consequentialism being impersonal and cruel, again virtue ethics is better situated. Dispensing with aggregate evaluations, it takes into account contextual considerations about AI developers, risk bearers and beneficiaries. Such contextual considerations encompass procedural as well as outcome-oriented aspects. This emphasis on context (especially through the virtue of practical wisdom) likewise places virtue ethics in a better position than consequentialism concerning the objection about the inaccuracy of formalised ways of thinking.

A welcome corollary is that, while avoiding these objections, virtue ethics does not exclude the costs or benefits of AI from being factored into developers' reasoning. That an ethical approach considers consequences does not necessarily mean that it is a consequentialist one[54]. The difference is that in consequentialism, which aims at the maximisation of good over bad consequences, comparisons between consequences are the sole or primary means of evaluation. In virtue ethics, which aims at a good, flourishing life more broadly, considerations of consequences are included among other, more important factors, particularly the agent's virtuous/vicious dispositions and reasoning. Instead of maximising for a single criterion, virtue ethics considers plural, heterogeneous and incommensurable values that are not salient in consequentialism. For example, it would not place AI developers' dispositions on the same scale, allowing the lack of one virtue to be compensated by the increased presence of another. In addition, a virtue ethics consideration of consequences would take into account the context of development/deployment, opposing predictions of AI bringing about certain

[54] Otherwise, almost any ethical theory would be «consequentialised». Conversely, any ethical theory that included some consideration of virtue would be converted into a virtue-ethical one.

consequences independently of social context or use. Such contextualisation allows virtue ethics to adapt to emerging technologies and different cultures in a way in which consequentialism cannot, despite being important to AI as a technology that crosses national, regional or cultural frontiers. Thus, virtue ethics preserves yet addresses the epistemic and moral uncertainty as well as complexity of the situation at hand, rather than reduce them to a simpler, fixed picture or mask them behind quantification as consequentialism does.

c. *Objections*

Although a virtue ethics approach to risk avoids the pitfalls of consequentialism, it can be criticised for failing the role expected of ethical theories, which is to provide a decision procedure, namely «an organized and systematic way of telling us what is the right thing to do»[55]. Just as technical manuals should do the intellectual heavy lifting for us, clarifying the steps we should follow to operate machinery, ethical theories should do the moral heavy lifting for us, issuing instructions we should follow to perform the morally right action in each circumstance. As the steps indicated – for machinery or right actions – are equally available to everyone, this «technical manual model» remains attractive, setting success standards for ethical theories[56].

Consequentialism abides by this model: «[i]t isolates one simple principle behind the directives of our everyday ethical discourse, and then tells us how to formulate this principle and apply it to tell us, systematically and specifically, what to do»[57]. Conversely, by not focusing on the traits of an action but the «qualities of agency» displayed therein, including the risk-taker's motivation, disposition, capacities and reasoning, virtue ethics struggles to identify in advance and in a manner applicable to everyone what a right action or morally acceptable risk would be[58]. It leaves agents without instructions precisely in morally fraught cases when identifying right action becomes critical[59].

[55] J. Annas, *Being Virtuous and Doing the Right Thing*, in «Proceedings and Addresses of the American Philosophical Association» 78, no. 2 (November 2004), p. 62, https://doi.org/10.2307/3219725. See also R.B. Louden, *On Some Vices of Virtue Ethics*, in «American Philosophical Quarterly» 21, no. 3 (July 1984), pp. 227-236.

[56] Annas, *op. cit*.

[57] Annas, *op. cit.*, p. 63.

[58] D. Cox, *Agent-Based Theories of Right Action*, in «Ethical Theory and Moral Practice» 9, no. 5 (October 2006), p. 506, https://doi.org/10.1007/s10677-006-9029-3.

[59] *Ibidem*.

Even if virtue ethics identified right-making features of an action, it is objected that deliberating along their lines would be impermissible[60]. Cox argues that if it is morally right to do x, it must be morally permissible to accurately deliberate about doing x[61]. Virtue ethics might consider that an act manifests the virtue of courage – and is thereby morally instead right – without allowing agents to explicitly deliberate performing said act *because* it manifests courage, lest they exhibit the vice of moral narcissism[62]. Breaking the link between performing and deliberating right action (specifically rendering the latter a violation of the former), the virtue-ethical theory of right action appears contradictory[63].

Two responses are plausible against this criticism. The first accepts that ethical theories should be action-guiding but questions whether virtue ethics fails to be so[64]. Virtue ethics suggests that an action is right if and only if it is what a virtuous agent would characteristically do in the circumstances[65]. Albeit considered under-specified, this theory of right action exhibits the same structure as consequentialism. For the latter, an action is right if and only if it promotes the best consequences, which is not action-guiding until one specifies what counts as the best consequences. Therefore, virtue ethics cannot be less action-guiding solely because it requires specification.

This theory of right action is additionally considered circular: it identifies the right action by reference to the virtuous agent, who, in turn, might be defined as one who performs right actions. Hence, one cannot know what a virtuous agent would do, unless one is already virtuous, in which case guidance is unnecessary. However, an agent may find and consult exemplars in their environment, a practice intuitively used yet unaccounted for in consequentialism. Alternatively, as this paper does, one may focus on canonical virtues/vices.

The second response challenges the very need for ethical theories to provide action guidance, in the form of a decision procedure, in order to be complete. The virtue-ethical theory of right action reproduces the manual

[60] *Ibidem*.

[61] *Ibidem*.

[62] *Ibidem*.

[63] Cox, *op. cit.*; J. Hacker-Wright, *Virtue Ethics without Right Action: Anscombe, Foot, and Contemporary Virtue Ethics*, in «The Journal of Value Inquiry» 44, no. 2 (March 2010), pp. 209-224, https://doi.org/10.1007/s10790-010-9218-0.

[64] R. Hursthouse, *On Virtue Ethics*, 1st ed., Oxford University Press, Oxford 1999; R. Hursthouse, *Normative Virtue Ethics*, in R. Crisp (ed)., *How Should One Live?*, Oxford University Press, Oxford 1996, pp. 19-33.

[65] Annas, *op. cit.*, p. 67.

model, albeit via a proxy, namely, the technical/virtuous expert, whose instructions are treated authoritatively[66]. However, the desirability of being «told what to do» by manuals or experts is questionable[67]. If what matters is merely the application of a theory that tells agents what to do rather than let them make their own moral decisions, praise and blame are more fitting to the theory itself than the agents' character, undermining the need for the latter's improvement[68].

Instead, virtue ethics offers a developmental and aspirational account. On this account, instructions and exemplars are starting points, but agents gradually develop an independent and critical understanding of what virtue requires, an understanding that might not only transcend but further oppose received learnings. Praise and blame are attributed to the agents' actions and decisions, which reflect their character, rather than the agents' application of a theory[69]. Hence, an all-purpose decision procedure available to anyone, regardless of their learning stage, background or character, would both be unrealistic and confine agents to the beginner's state[70]. Contrary to following instructions, agents must do the moral heavy lifting on their own.

Returning to Cox's objection, virtuous agents thus do not ask how an act would reflect on their character but how «experienced people of good character» would act in these circumstances[71]. These people are not necessarily fully virtuous but are better than us (more generous, temperate, and so on). As discussed in sub-section 4.b, such deliberation encompasses the consequences of one's actions on others, rendering the charge of moral narcissism void[72].

Overall, even if virtue ethics is not deemed sufficiently action-guiding, it offers more important guidance on how to improve one's reasoning, considering where agents themselves and their role models stand in their moral development. It recognises that «moral life is not static; it is always developing. When it comes to working out the right thing to do, we cannot shift the work to a theory, however excellent, because we, unlike the theories, are always learning, and so we are always aspiring to do better»[73]. This aspirational, developmental approach of virtue ethics renders it apt for grappling with risk.

66 Annas, *op. cit*., p. 68.
67 Annas, *op. cit.*, pp. 64-65.
68 Annas, *op. cit.*, p. 65.
69 Annas, *op. cit*.
70 Annas, *op. cit*.
71 Hacker-Wright, *op. cit.*, p. 220.
72 Hacker-Wright, *op. cit*.
73 Annas, *op. cit*, p. 74.

d. *Virtuous AI risk-takers*

By focusing on agents, the suggested approach moves from an analysis of *risk* as a noun to an analysis of *(to) risk* as a verb[74]. The morality of risking, then, depends on the risk-taker's dispositions and responsiveness to contextual features of the situation[75]. While several virtuous dispositions are supported in the literature, four are considered 'cardinal' by philosophers in antiquity and later: courage, temperance, justice and (practical) wisdom. Next to these, a fifth one, friendship, is key in Aristotelian virtue ethics.

Whereas thin concepts (e.g., right/wrong, good/bad) denote evaluation only, virtues are thick concepts that combine evaluative with non-evaluative descriptions and thereby are more information-rich. The content of virtues can be specified to bring out dimensions appropriate or important to each context. As part of such a specification effort, and adopting the tenets of Aristotle's virtue ethics as these are fleshed out in his *Nicomachean Ethics*, the following questions are suggested for AI risk-takers' self-assessment[76].

Courage

Courage (*andreia*) is a mean state between fear and over-confidence, distinguished by its motivation. Taking risks to avoid another evil (e.g., repercussions or reproach) indicates cowardice, whereas courage stems from a motivation to achieve what is noble and good[77]. Questions to consider include:

– Do you strive to strike a balance between risk-averse and risk-seeking behaviour?
– Are you disposed to put yourself in harm's way (e.g., to confront internal/external pressures) to promote users' flourishing? Are you disposed to speak up about errors, limitations or blind spots, whether yours or those of the AI system?
– Are you disposed to embrace external criticism, divergent scientific views and other sources of knowledge/expertise to develop scientifically excellent systems?

[74] Hansson, *Philosophical Perspectives on Risk*, cit., p. 30.

[75] Athanassoulis, Ross, *op. cit*. I interpret these two conditions as conjunctively (rather than disjunctively) required.

[76] Aristotle, *Nicomachean Ethics*, in J. Barnes (ed.), *The Complete Works of Aristotle: The Revised Oxford Translation*, trans. W.D. Ross and J.O. Urmson, vol. 2, Bollingen, 71: 2, Princeton University Press, Princeton 1985, pp. 1729-1867. Hereafter, *NE*, with references in Bekker numbering.

[77] *NE*, III.7, 1116a10-15; NE, III.8, 1116a25-30.

Temperance

Temperance (*sophrosune*) regulates pleasure[78]. Intemperate agents take pleasure in things that are wrong or take pleasure in things that are right but do so in a wrong way[79]. Temperate agents seek pleasures that promote health, well-being or nobility and do not exceed the available means. Questions to consider include:

– Are you disposed to forgo pleasurable returns (e.g., economic rewards) or self-indulgent goods when deciding about AI development/deployment?
– Are you disposed to develop systems that promote users' health, safety and overall good lives?
– Are you disposed to strike a sustainable balance in terms of the (environmental) resources used as a means to AI development/deployment?

Justice

Justice (*dikaiosune*) describes lawful and equal agents, whereas unjust agents are greedy, unfair and unlawful[80]. By benefitting those interacting with the just agent (e.g., fellow citizens), justice is strongly relational[81]. It comprises two types: distributive justice concerns the distribution of honour, wealth or anything shared among citizens; corrective justice concerns the correction of voluntary (e.g., commercial) or involuntary (e.g., mandatory) relations/transactions[82]. Questions to consider include:

– How will AI risks and benefits be shared among users (present and future ones)? Which are the different stakeholders and what are their particular status and needs?
– How voluntary or involuntary will the acceptance of the AI system and its risks be by users? Are these risks self- or other-imposed?
– How will you correct for possible harms? Are you open to taking responsibility for this AI system? Have you established chains of accountability?

Friendship

Friendship (*philia*) requires that (i) parties should express mutual goodwill and (ii) this goodwill should stem from a pursuit of the noble, pleasant or useful, with (ii) determining the kind of friendship and content of the good

78 *NE*, III.10, 1117b25-30.
79 *NE*, III.11, 1118b22-27.
80 *NE*, V.1, 1129a31-1129b1.
81 *NE*, V.1, 1130a2-5.
82 *NE*, V.2, 1130b30-1131a9.

that parties wish for each other[83]. Friendships also vary on the basis of association (*koinonia*) among the parties[84]. All joint undertakings foster friendship, with the broadest one being political association between citizens and accordingly political/civic friendship[85]. Questions to consider include:

– Do your decisions to risk demonstrate goodwill (e.g., care and empathy) towards users?
– Is the decision to risk undertaken jointly with other stakeholders? Are you disposed to engage non-experts, particularly citizens possibly affected by the AI system, in the decision-making process?
– Does the decision to risk serve a mutual pursuit of a noble, pleasant or useful objective?

Practical wisdom

Practical wisdom (*phronesis*) is an intellectual virtue that shapes the aforementioned moral ones[86]. It is a practical disposition that involves right reasoning about what is good or bad for humans[87]. Its practicality means that it does not focus on theoretical or abstract goods but on the actions that bring about the practical or moral good[88]. As such, it concurrently assesses the means and ends of a particular action. Questions to consider include:

– What are the broader end(s) that risk-taking does or should serve in this case?
– What are the most suitable means to achieve these ends? Are there less risky means available?
– What are the morally salient features of this situation? What risks does the AI system pose (e.g., on users' health, safety, social and psychological states, rights)? Have you attempted to imagine how these risks might be perceived from the perspective of users?

Briefly put, in AI-related decision-making, developers should aspire to take the risks that a courageous, temperate, just, friendship-promoting, and practically wise agent would accept. As stable dispositions, such virtues require practice to become part of one's way of life. This self-assessment should be iteratively performed throughout the AI lifecycle, while the risk-

[83] *NE*, VIII.2-VIII.3.
[84] *NE*, VIII.9, 1159b25-32.
[85] *NE*, VIII.9, 1160a9-14; NE, IX.6, 1167b1-5.
[86] *NE*, VI.3, 1139b15-17; NE, VI.13, 1144b30-32.
[87] *NE*, VI.5, 1140b1-5.
[88] *NE*, VI.7, 1141b10-15.

taking in which AI developers habitually engage over time is of greater interest than is risk-taking in extreme or high-profile instances. Reliable access to training, role models and virtue-friendly environments is thereby necessary for developers' ongoing self-cultivation.

Moreover, unlike the consequentialist approach concentrating on agents' actions at the expense of their motives, virtues are dispositions to act for the right reasons, implying that AI developers should justify their answers to these self-assessment questions. Without an understanding of the developers' reasoning, third parties can neither evaluate whether developers are virtuous risk-takers nor hold them responsible or trust them[89]. However, virtues are also dispositions to act with the right emotional responses. Illuminating the relevance of emotions to decision-making, virtue ethics challenges conventional portrayals and ideals technologists as purely analytical, impassive professionals.

Although none of these questions determines on its own the acceptability of AI risk, those highlighting considerations of means and ends are particularly important, albeit often neglected. Developers should justify the riskiness of their AI system against the backdrop of not only other systems or digital solutions but also non-technical options. Doing so will counter the «entrenched assumption that the mere advancement to market of a new product, process or technology is demonstration of social "benefit"»[90]. Additionally, comparing AI systems with both technical and non-technical means will illuminate whether their adoption is voluntarily chosen among multiple other options, merely the best among a limited range of alternative options, or even the sole option suitable for achieving the desired end.

Overall, virtue ethics provides a heuristic that does not face the same challenges as consequentialism, as it does not depend on outcomes, but manages to capture more of the ethically relevant dimensions of risk-taking, furnishing a broader *and* more tailored viewpoint. This set of questions is put forward as a preliminary framework. Far from painting a complete picture, they can be supplemented with other virtues or altogether different considerations[91]; still, they may extend the range of normative questions factored into developers' decision-making and serve as ideals to which developers may aspire.

[89] A. Ross, N. Athanassoulis, *Risk and Virtue Ethics*, in S. Roeser *et al.* (ed.), *Handbook of Risk Theory*, Springer Netherlands, Dordrecht 2012, pp. 833-856, https://doi.org/10.1007/978-94-007-1433-5_33; Athanassoulis, Ross, *op. cit*.

[90] Felt *et al.*, *op. cit.*, p. 84.

[91] See, e.g., S. Vallor, *Technology and the Virtues: A Philosophical Guide to a Future Worth Wanting*, Oxford University Press, New York 2016, https://doi.org/10.1093/acprof:oso/9780190498511.001.0001.

5. *Where to now?*

While AI risk appears like a novel threat to specific aspects or the entirety of human life, concerns about risk have permeated the societal and legal fabric over the past years. Within this «enhanced risk apparatus» of society in general and technology regulation in particular, risk is largely approached in an objectivist, technical manner[92]. Instead, this paper highlights its normative dimensions and illustrates that technocratic approaches to its governance are incomplete or even inaccurate without normative, particularly virtue-ethical, ones.

My illustration has not been exhaustive, especially since the contextual nature of virtue ethics resists codification; it does, however, lay the foundations for further fundamental and applied research. Future applications of the virtue-ethical approach could tailor the (cardinal or other) virtues to risks or challenging situations identified in the AI development literature. It would also be promising to examine whether these virtues could be exercised not only by AI developers as individual risk-takers but in the form of group or institutional virtues exercised by organisations as collective risk-takers. Policymakers are likewise urged to acknowledge the normative dimensions of risk and integrate them into risk-based regulation. Such dimensions may accordingly be embedded in efforts to audit AI.

At the same time, considerations of risk refine ethical reasoning itself, as in practice we operate in far less certain environments than the ones assumed by consequentialism. This uncertainty and its nuances are better captured by virtue ethics. While risk-takers cannot guarantee that the consequences of their actions will eventually occur, they have greater control over the quality of the decision-making that results in said consequences, and this distinction bears on ascriptions of responsibility. This is why virtue ethics attributes primacy to risk-takers' attitudes and their attuned responsiveness to context rather than cost-benefit calculations. To grossly simplify, the goodness (or not) of risk is deduced from the goodness (or not) of one's character. As such, virtue ethics is applicable to risks posed by emerging technologies, including AI, and the ever-changing context that these shape. The aim is not that the actions performed or systems built by AI developers be faultless or that their benefits score higher than their costs in relevant calculations but that AI developers prove to be the courageous, temperate, just, friendship-promoting, and practically wise risk-takers that our society

[92] van Dijk, Gellert, Rommetveit, *op. cit.*, p. 288.

needs. This might prove a more realistic aim, dispensing with modern illusory perceptions of absolute risk objectivity and controllability.

AI practitioners broadly speaking could operationalise the virtue-ethical framework proposed herein as a complementary to or integral facet of AI risk governance, for instance, through the integration of its questions into their codes of conduct. Additionally, policymakers/legislators, employers/managers, and educators are urged to foster organisational cultures and broader environments conducive to the development and exercise of virtues by AI developers. Virtue ethics may thus serve as a compass for navigating AI-induced risk and discerning the moral needs of our messy world writ large.

Acknowledgements

I am grateful for valuable comments from an anonymous reviewer. The research for this paper has been funded through a PhD Fellowship for Fundamental Research by the Research Foundation Flanders (Fonds Wetenschappelijk Onderzoek), grant no. 1151621N/1151623N.

Abstract

Risk increasingly permeates technology regulation, as exemplified by the EU's General Data Protection Regulation and Artificial Intelligence (AI) Act. Nonetheless, contrary to common distinctions between an objective stage of risk assessment and a normative one of risk management, I argue that risk governance is normative throughout; hence, it should accordingly integrate normative considerations. To achieve this integration, this paper adopts a normative perspective on AI risk governance in particular. It examines AI-induced risk from the dominant approach of consequentialism, highlighting its limitations in conditions of uncertainty. It suggests virtue ethics as an alternative yet overlooked approach to AI-induced risk and concludes with implications of this approach for research, policy and practice.

Keywords: Virtue ethics; risk-based approach; Artificial Intelligence; AI ethics; AI risk.

Anastasia Siapka
Centre for IT & IP Law (CiTiP), KU Leuven
anastasia.siapka@kuleuven.be