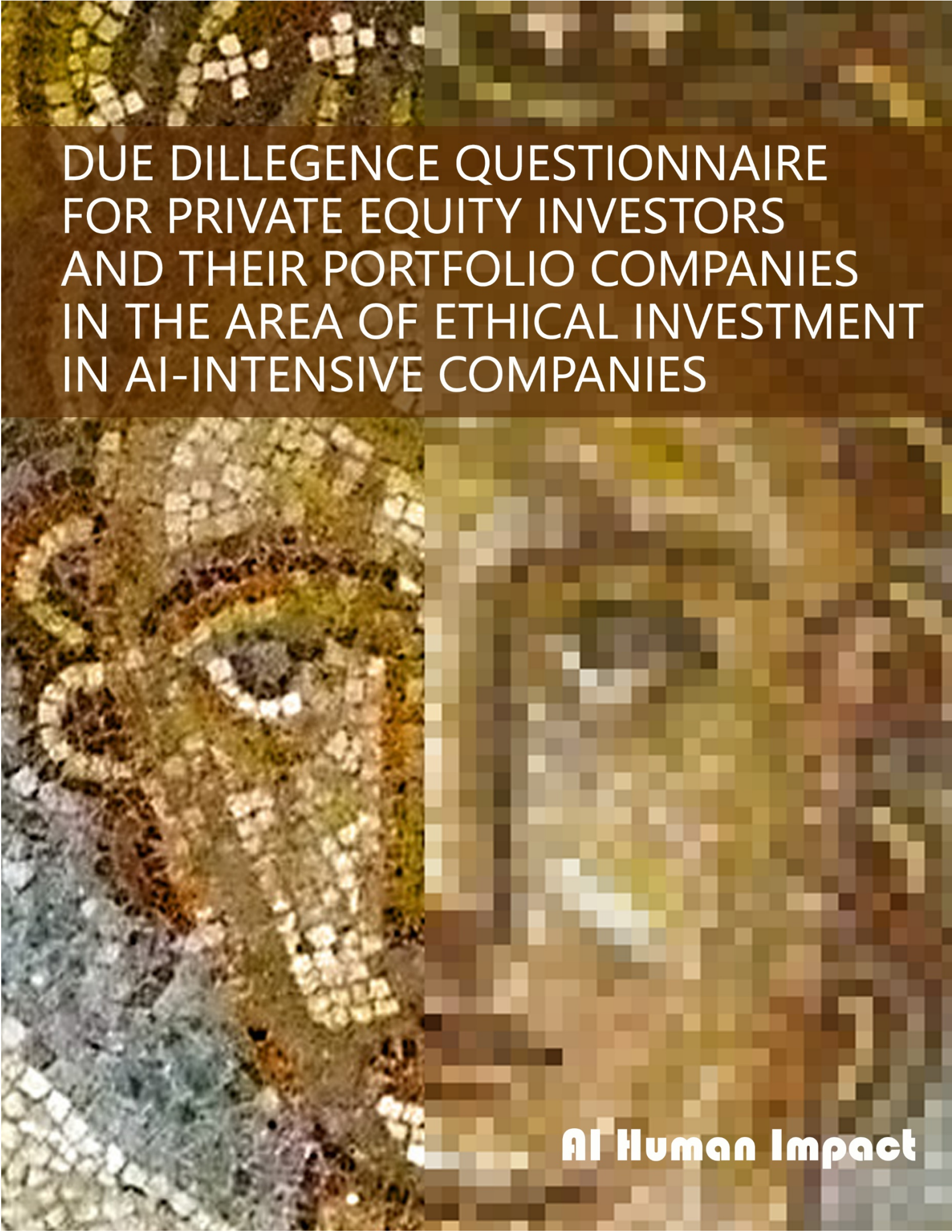


DUE DILLEGENCE QUESTIONNAIRE FOR PRIVATE EQUITY INVESTORS AND THEIR PORTFOLIO COMPANIES IN THE AREA OF ETHICAL INVESTMENT IN AI-INTENSIVE COMPANIES

AI Human Impact

Acknowledgements
Contributions to this document gratefully acknowledged:

AI Human Impact, 164 Lafayette Avenue, STE 1, Brooklyn NY, USA
T +1 949 677 0526
www.AIHumanImpact.Fund connect@aihumanimpact.fund

Venus, Roman floor mosaic from the Bignor Roman Villa, Sussex
Venus at her Toilet with Mirror, from Thuburbo Majus, Bardo Museum, Tunis

INTENTION

This questionnaire helps AI-intensive companies define and demonstrate their ethical commitments at the intersection of technology and humanity.

Some commitments are absolute and may be scored hierarchically. Companies that dutifully protect users' personal information from malicious hackers are superior to those that do not, assuming all else is equal.

Most commitments are relative, and best understood as tensions. For instance, users and subjects of AI reasonably want to preserve control over their personal information, but effective AI services in insurance, healthcare and elsewhere can require personal exposure and widespread data sharing. Ethics here is not about right and wrong so much as understanding and prioritizing values: Between privacy and AI performance, which prevails? Because there is no right answer, the intention is not to rank but distinguish companies in terms of their ethical orientations.

Surveillance capitalism originates in companies staking a claim to people's lives as behavioral data. There follows a sharp rise in inequality between what I can know and what can be known about me. In the real world this splinters shared reality, poisons social discourse, paralyzes democratic politics and sometimes instigates violence and death.

Shoshana Zuboff
New York Times

This is why surveillance capitalism has boomed. Like scores of others, I've decided that I'm OK with giving up personal data in order to keep getting convenient, cheap (or free) services. Despite the known episodes of firms misusing data, the ease and quality of life under the reign of Big Tech generally seems worth it.

Erica Pandey
Axios

Motivation

For companies, the goal of an ethical profile is to catalyze more and faster artificial intelligence innovation in three ways. First scenes of potential social rejection are illuminated, including unfair outcomes resulting from inadequate training data. More significantly, irresolvable dilemmas are delineated so that they can be capably navigated, such as the facial recognition trade-off between the right to privacy and the desire for quick conveniences. Finally, an ethical evaluation identifies AI engineering that expands human capabilities and potential, and so orients development toward sustainable gains.

Companies catalyze more and faster AI innovation

For investors, the leading benefit is increased ability to align equity with their own values: investor freedom expands on the ethical level. AI ethical profiling also provides nonfinancial information that contributes to a portfolio's risk-adjusted return.

Investors align equity with their own values

Process

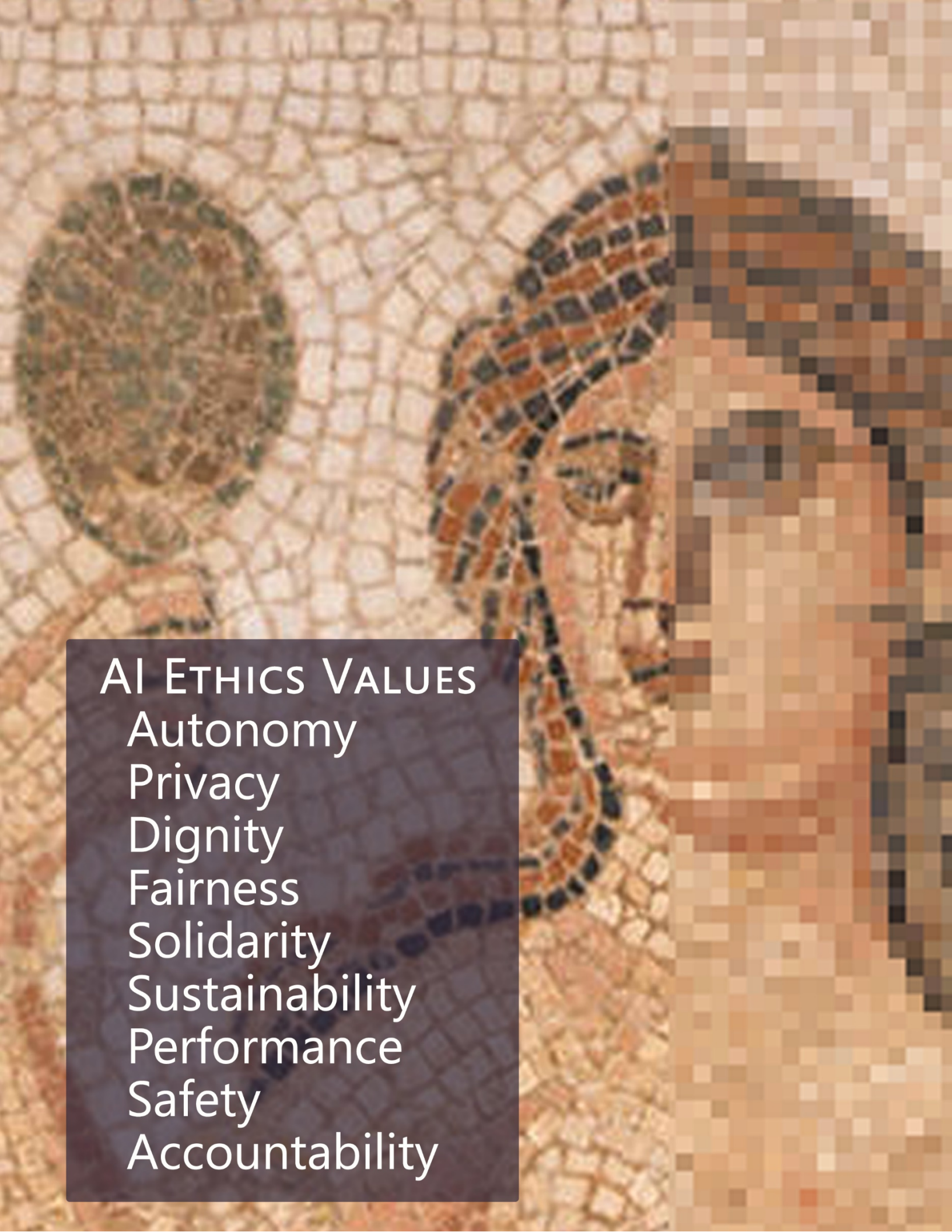
Initially, each of nine principles of AI ethics organizes specific questions designed to circumscribe the company's profile in that specific area. Not every question applies to every company.

Subsequently, respondents are asked about tensions and tradeoffs between the values, and the final question summarizes the ethical profile by asking for a ranking of the discussed values: When choices are unavoidable, which way does the company lean? There are no preferred responses, but the differences allow investors to act in accord with their own values.

Notes

For a fuller explanation of the principles' origin, please see *AI human impact: toward a model for ethical investing in AI-intensive companies* in the *Journal of Sustainable Finance & Investment*, or visit AIHumanImpact.Fund.

Some the questions in this document have been modified from the European Council's Assessment List for Trustworthy Artificial Intelligence available here: https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment.


AI Ethics Values
Autonomy
Privacy
Dignity
Fairness
Solidarity
Sustainability
Performance
Safety
Accountability

Autonomy
Self-determination

1. Does the AI empower users to do new things?
    - Does it provide opportunities that were previously unthinkable?
    - Does it provide opportunities that were previously unavailable?

2. Does the AI short-circuit human self-determination?
    - Have measures been taken to mitigate the risk of manipulation, and to eliminate dark patterns?
    - Has the risk of addiction been minimized?
    - If the AI system could generate over-reliance by end-users, are procedures in place to avoid end-user over-reliance?

3. Could the AI system affect human autonomy by interfering with the end-user's decision-making process in any other unintended and undesirable way?

Dignity

Users hold intrinsic value: they are subjects or ends in themselves, not objects, means, tools, instruments

1. Are users respected as the *reason* for the AI? Does the machine primarily serve the users' projects and goals, or does it employ users as tools or instruments in external projects?

2. Are mechanisms established to inform users about the full range of purposes, and the limitations of the decisions generated by the AI system?
    - Are the technical limitations and potential risks of the AI communicated to users, such as its level of accuracy and/ or error rates?

3. Does the AI system simulate social interaction with or between end-users or subjects (chatbots, robo-lawyers and similar)?
    - Could the AI system generate confusion about whether users are interacting with a human or AI system? Are end-users informed that they are interacting with an AI system?

Privacy
Control over access to one's own personal information

1. Is the AI system trained or developed by using or processing personal data?

2. Do users maintain control over access to their personal information? Is it within their power to conceal and to reveal what is known and shared?

3. Is data minimization, in particular personal data, in effect?

4. Is the AI system aligned with relevant standards (e.g., ISO, IEEE) or widely adopted protocols for (daily) data management and governance?

5. Are the following measures, or non-European equivalents, established?
    - Data Protection Impact Assessment (DPIA).
    - Designate a Data Protection Officer (DPO) and include them at an early state in the development, procurement or use phase of the AI system.
    - Oversight mechanisms for data processing (including limiting access to qualified personnel, mechanisms for logging data access, and for making modifications).
    - Measures to achieve privacy-by-design and default (e.g., encryption, pseudonymization, aggregation, anonymization).

- The right to withdraw consent, the right to object, and the right to be forgotten implemented into the AI's development.

6. Have privacy and data protection implications been considered for data collected, generated, or processed over the course of the AI system's life cycle?

Fairness

Equals treated equally and unequals treated proportionately unequally (Aristotle)

1. Are equals treated equally and unequals treated unequally by the AI? (Aristotle's definition of fairness.)

2. Have procedures been established to avoid creating or reinforcing unfair bias in the AI system for input data, as well as for the algorithm design?

3. Is your statistical definition of fairness commonly used? Were other definitions of fairness considered?
    - Was a quantitative analysis or metric developed to test the applied definition of fairness?

4. Was the diversity and representativeness of end-users and subjects in the data considered?
    - Were tests applied for specific target groups, or problematic use cases?
    - Did you consult with the impacted communities about fairness definitions, for example representatives of the elderly, or persons with disabilities?
    - Were publicly available technical tools that are state-of-the-art researched to improve understanding of the data, model, and performance?
    - Did you assess and put in place processes to test and monitor for potential biases during the entire lifecycle of the AI system (e.g., biases due to possible limitations stemming from the composition of the used data sets (lack of diversity, non-representativeness)?

5. Is there a mechanism for flagging issues related to bias, discrimination, or poor performance of the AI?
    - Are clear steps and ways of communicating established for how and to whom such issues can be raised?

6. In addition to the users and subjects, are subjects that could potentially be affected by the AI system identified?

Solidarity

No one left behind: AI most benefits those who have least (John Rawls)

1. Is the AI designed so that no one is left behind?

2. Does the AI deliver the maximum advantage to those users who are most disadvantaged? (Does the most go to those who have least? John Rawls' definition of Solidarity/Justice.)

3. Is the AI adequate to the variety of preferences and abilities in society?
    - Were mechanisms considered to include the participation of the widest possible range of stakeholders in the AI's design and development?

4. Were Universal Design principles taken into account during every step of the planning and development?
    - Did you assess whether the AI system's user interface is usable by those with special needs or disabilities, or those at risk of exclusion?
    - Were end-users or subjects in need for assistive technology consulted during the planning and development phase of the AI system?

5. Was the impact of the AI on all potential subjects taken into account?

6. Could there be groups who might be disproportionately affected by the outcomes of the AI system?

Sustainability
Enduring social flourishing

1. For societies around the world, does the AI advance toward, or recede from the United Nations' Sustainable Development Goals? Please elaborate for each relevant goal. For elaboration of the particular goals, see: https://sdgs.un.org/goals

    1: No poverty
    2: Zero hunger
    3: Good health and well-being
    4: Quality education
    5: Gender equality
    6: Clean water and sanitation
    7: Affordable and clean energy
    8: Decent work and economic growth
    9: Resilient industry, innovation, and infrastructure
    10: Reducing inequalities
    11: Sustainable cities and communities
    12: Responsible consumption and production
    13: Climate change
    14: Life below water
    15: Life on land

Performance
The machine works

1. Has a definition of what counts as performance been articulated for the AI: Accuracy? Rapidity? Efficiency? Convenience? Pleasure?

2. Is there a clear and distinct performance metric?
    - Does the metric correspond with human experience?

3. Does the AI outperform humans? Other machines?

Safety
Robust, resilient against dangers

1. Could the AI adversely affect human or societal safety in case of risks or threats such as design or technical faults, defects, outages, attacks, misuse, inappropriate or malicious use?
    - Were the possible threats to the AI system identified (design faults, technical faults, environmental threats), and also the possible consequences?
    - Is there a process to continuously measure and assess risks?
    - Is there a proper procedure for handling the cases where the AI system yields results with a low confidence score?
    - Were potential negative consequences from the AI system learning novel or unusual methods to score well on its objective function considered?
    - Can the AI system's operation invalidate the data or assumptions it was trained on? Could this lead to adverse effects?

2. Could the AI system cause critical, adversarial, or damaging consequences in case of low reliability or reproducibility?
    - Are there verification and validation methods, and documentation (e.g., logging) to evaluate and ensure different aspects of the AI system’s reliability and reproducibility?

3. Are failsafe fallback plans defined and tested to address AI system errors of whatever origin, and are governance procedures in place to trigger them?

4. Have the humans in-the-loop (human intervention in every decision of the system), on-the-loop (human monitoring and potential intervention in the system's operation), in-command (human overseeing the overall activity of the AI system, including its broader economic, societal, legal and ethical impact, and the ability to decide when and how to use the AI system in any particular situation, including the decision not to use an AI system in a particular situation, and the ability to override a decision made by an AI system) been given specific training on how to exercise oversight?
    - Is there a 'stop button' or procedure to safely abort an operation when needed?

5. How exposed is the AI system to cyber-attacks?
    - Were different types of vulnerabilities and potential entry points for attacks considered, such as:
        - Data poisoning (i.e., manipulation of training data).
        - Model evasion (i.e., classifying the data according to the attacker's will).
        - Model inversion (i.e., infer the model parameters)
    - Did you red-team and/or penetration test the system?

6. Is the AI system certified for cybersecurity and compliant with applicable security standards?

Accountability
Responsibility for AI processing is attributable

1. Would erroneous or otherwise inaccurate output significantly affect human life?

2. When the AI goes right or wrong, can credit or blame be accurately assigned? Can responsibility for the development, deployment, use and output of AI systems be attributed?
    - Can you explain the AI's decisions to users? Are you transparent about the limitations of explanations?
    - Can you trace back to which data, AI model, and/or rules were used by the AI system to make a decision or recommendation?

3. Are accessible mechanisms for accountability in place to ensure contestability and/or redress when adverse or unjust impacts occur?
    - Have redress by design mechanisms been put in place?

4. Did you establish mechanisms that facilitate the AI system's auditability (e.g., traceability of the development process, the sourcing of training data and the logging of the AI system's processes, outcomes, positive and negative impact)?
    - Did you consider establishing an AI ethics review board or a similar mechanism to discuss the overall accountability and ethics practices?
    - Does review go beyond the development phase?

5. Did you establish a process for third parties (e.g., suppliers, end-users, subjects, distributors/vendors or workers) to report potential vulnerabilities, risks, or biases in the AI system?

Tensions & Tradeoffs Between AI Ethics Values

The values of AI ethics break naturally into three categories founded on the history of philosophy and ethics.

- Individual ethics tends to correlate with a libertarian view focusing on individuals and their opportunities before society and its wellbeing. (Though flourishing individuals may lead to social wellbeing.)

- Social ethics tends to correlate with a communitarian or utilitarian world view focusing on society and its wellbeing before individuals and their opportunities or flourishing. (Though social wellbeing may lead to individuals flourishing.)

- Technological ethics corresponds with an aesthetic view of computer science and engineering, one where innovation and discovery is valuable in itself – like art – independent of the human consequences.

Tensions and tradeoffs exist between these categories. For example, AI finance empowers individuals to participate competitively in marketplaces previously reserved for large companies, but it also jeopardizes social cohesion and sustainability by facilitating the accumulation of outsized personal wealth. Contrarily, social credit scoring like that implemented in China potentially wraps every aspect of personal

lives into a larger project for social harmony, but also jeopardizes individual freedom and privacy.

Similarly, tensions exist within the categories. By providing health alerts and recommendations, medical wearables can increase independence and autonomy, but the privacy cost may be significant.

There are no intrinsically right or wrong responses to these ethical tensions, but companies will have their own priorities. This question asks that you rank the values of AI ethics in order from 1 to 9, with 1 being the highest priority in cases where decisions must be made, and 9 being the lowest.

_____Autonomy

_____Privacy

_____Dignity

_____Fairness

_____Solidarity

_____Sustainability

_____Performance

_____Safety

_____Accountability

End