Wachter, S., Mittelstadt, B., & Floridi, L. (2017). Transparent, explainable, and accountable AI for robotics. Science Robotics, 2(6), eaan6080.

# Transparent, Explainable, and Accountable AI for Robotics

Sandra Wachter,[1,2] Brent Mittelstadt,[2,3,1] Luciano Floridi[1,2]

[1]Oxford Internet Institute, University of Oxford, 1 St Giles, Oxford, OX1 3JS, United Kingdom;
[2]The Alan Turing Institute, British Library, 96 Euston Rd, London, NW1 2DB, United Kingdom;
[3]Department of Science and Technology Studies, University College London, 22 Gordon Square, London, WC1E 6BT, United Kingdom.

**Correspondence:** Sandra Wachter, sandra.wachter@oii.ox.ac.uk

Recent governmental statements from the United States (USA) (1, 2), the European Union (EU) (3), and China (4) identify artificial intelligence (AI) and robotics as economic and policy priorities. Despite this enthusiasm, challenges remain. Systems can make unfair and discriminatory decisions, replicate or develop biases, and behave in inscrutable and unexpected ways in highly sensitive environments that put human interests and safety at risk (5). For example, Tesla's self-driving cars, policing robot Knightscope, or companion robot Pepper autonomously decides whether something is a pedestrian or another car, whether an individual poses a threat, or which emotion(s) the user is experiencing. In response, pressure is mounting to make algorithms, AI, and robotics fair, trans parent, explainable, and therefore accountable.

These challenges have been reflected in regulation applicable to automated systems since the 1970s. In the USA, the Equal Credit Opportunity Act (ECOA) and the Fair Credit Reporting Act (FCRA) aim to increase transparency in the credit industry (6) and indirectly affect automated systems. Consumers are guaranteed notifications of reasons for adverse actions, including those based on automated scoring systems. More directly, in the EU, the 1995 Data Protection Directive guarantees individuals a "right of access" to demand "know ledge of the logic involved" in automated decision-making, for example, about creditworthiness.

Since the 1970s, algorithmic systems and their accountability issues have grown in scale and complexity. American and European policies now appear to be diverging on how to close current accountability gaps in AI. In the USA, notifications guaranteed by the ECOA and FCRA remain. However, recent recommendations on AI focus more on ethical design, education, and self-regulation than on individual rights (1, 2). In comparison, the EU continues exploring a "hard" regulatory approach with legally enforceable rights. This divergence may reflect the new complexity of regulating AI and robotics compared to previous automated systems. The inscrutability and the diversity of AI complicate the legal codification of rights, which, if too broad or narrow, can inadvertently hamper innovation or provide little meaningful protection.

This tension can be seen in recent European policy debate on the General Data Protection

Regulation (GDPR) and the European Parliament's resolution on "Civil Law Rules on Robotics" (3). One potential accountability mechanism has received great attention: the GDPR's "right to explanation." This would be robust but potentially disruptive and technically challenging for AI, requiring certain automated decisions to be explained to individuals. Despite a proposal by the European Parliament to guarantee a "right to explanation," this appears only in a nonbinding Recital (7). Elsewhere, individuals are guaranteed "meaningful information" about the "logic involved" in certain automated decision making through the GDPR's "right of access." Although the Regulation fails to define the scope of information to be provided in practice, only a general, easily understood overview of system functionality is likely to be required (7).

The civil law resolution on robotics similarly struggles to define precise accountability mechanisms. Transparency tools to explain the "rationale" and "logic" of robotic behavior and decision-making aided by AI are called for but left undefined (3). The Parliament's Committee on Civil Liberties, Justice, and Home Affairs called for compliance with the GDPR in future civil law addressing robotics (8). Several data protection safeguards were explicitly highlighted, including "the right to obtain an explanation" and "information obligations" (e.g., the right of access). Although GDPR compliance is still called for, both safeguards are no longer explicitly mentioned in the final resolution (3). European legislators thus missed a second opportunity to clarify the GDPR's accountability requirements for AI and robotics.

Issues remain, even if future civil law rules for robotics are fully compliant with the GDPR's safeguards against automated decision making. The safeguards only apply to decisions "based solely on automated processing," which may exclude many robotic systems (9). There is reluctance in high-risk areas (e.g., transport) to remove humans entirely from the loop. The outcome may be that robotic decision making would not qualify as "solely" automated. Ironically, this reluctance could make systems less accountable by preventing the GDPR's safeguards from applying. Automated decisions must also have "legal" or "significant" effects for safeguards to apply (Fig. 1), although a definition of such effects is not provided. Only two examples are given: online credit applications and e-recruiting. It remains to be seen whether autonomous robotic behaviors will have "legal" or "significant" effects and how levels of autonomy will influence this definition (9).

Designing imprecise regulation that treats decision-making algorithms, AI, and robotics separately is dangerous. It misinterprets their legal and ethical challenges as unrelated. Concerns about fairness, transparency, interpretability, and accountability are equivalent, have the same genesis, and must be addressed together, regardless of the mix of hardware, software, and data involved. For example, security robots and predictive policing software identify threats with the

same method (automated processing) and purpose (public safety). Hence, the desire to understand both systems is the same.

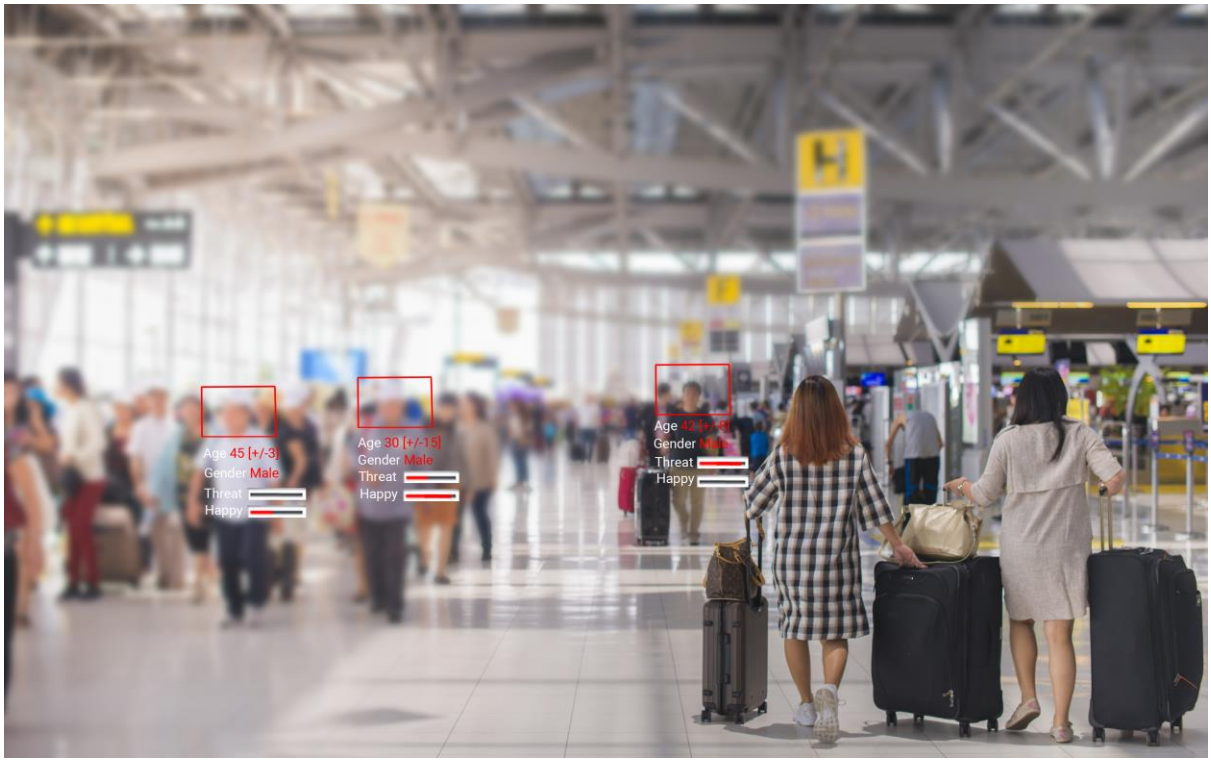


*Figure 1 - Security or companion robots detecting threat level or mood solely based on automated processing could produce "significant" effects for an individual, but it remains unclear whether such robotic decisions fall within the scope of the GDPR's safeguards. Photo credit: Shutterstock/Anucha Maneechote. Design: Adham Tamer, Oxford Internet Institute*

These issues will only grow in importance. Beijing will soon issue a national development plan for AI (10). It will be interesting to see whether China addresses AI's accountability challenges and, if so, adopts a self-regulatory or "hard law" approach comparable to the USA or EU. Other mechanisms may also be expanded, such as pre-deployment software certification schemes required by China's Cybersecurity Law.

Regulatory and technical accountability mechanisms will be effective only if designed by taking into account the common functionality and diverse complexity of algorithms, AI, and robotics. Several considerations require further research:

**How can human-interpretable systems be designed without sacrificing performance?** Interpretability is often perceived to be at odds with model accuracy and efficiency in machine learning. In robotics, methods are needed to provide legally required explanations without significantly hampering performance, for example, using proxy or simplified models or rule extraction.

**How can transparency and accountability be achieved in inscrutable systems?** Inscrutability in AI challenges calls for transparency. Mechanisms not reliant on full interpretability, including pre-deployment certification and algorithmic auditing (5), require further development to ensure transparency and accountability in opaque systems. It remains to be seen whether such "black box" approaches that assess inputs and outputs will comply with legal requirements.

**How can parallels between emerging systems be identified to set accountability requirements?** Regulatory standards need to be developed to set system- and context-dependent accountability requirements based on potential biased and discriminatory decision-making and risks to safety, fairness, and privacy.

## References


1. National Science and Technology Council, "Preparing for the future of artificial intelligence" (Executive Office of the President, 2016), (available at https://www.whitehouse.gov/sites/default/files/whitehouse_files/microsites/ostp/NSTC/preparing_for_the_future_of_ai.pdf).

2. National Science and Technology Council, "The National Artificial Intelligence Research and Development Strategic Plan" (Executive Office of the President, 2016), (available at https://www.nitrd.gov/PUBS/national_ai_rd_strategic_plan.pdf).

3. European Parliament, "Civil Law Rules on Robotics - European Parliament resolution of 16 February 2017 with recommendations to the Commission on Civil Law Rules on Robotics (2015/2103(INL))" (P8_TA-PROV(2017)00 51, European Parliament, 2017), (available at http://www.europarl.europa.eu/sides/getDoc.do?pubRef=-//EP//NONSGML+TA+P8-TA-2017-0051+0+DOC+PDF+V0//EN).

4. X. Bo, China rolls out three-year program for AI growth. *Xinhua News* (2016), (available at http://news.xinhuanet.com/english/2016-05/23/c_135382029.htm).

5. B. Mittelstadt, P. Allo, M. Taddeo, S. Wachter, L. Floridi, The ethics of algorithms: Mapping the debate. *Big Data Soc.* **3** (2016), doi:10.1177/2053951716679679.

6. W. F. Taylor, Meeting the equal credit opportunity act's specificity requirement: Judgmental and statistical scoring systems. *Buffalo Law Rev.* **29**, 73-130 (1980).

7. S. Wachter, B. Mittelstadt, L. Floridi, Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation. *Int. Data Priv. Law* (2017) (available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2903469).

8. European Parliament Committee on Legal Affairs, "Report with recommendations to the Commission on Civil Law Rules on Robotics" (2015/2103(INL), European Parliament, 2017), (available at http://www.europarl.europa.eu/sides/getDoc.do?pubRef=-//EP//NONSGML+REPORT+A8-2017-0005+0+DOC+PDF+V0//EN).

9. G.-Z. Yang *et al.*, Medical robotics—Regulatory, ethical, and legal considerations for

increasing levels of autonomy. *Sci. Robot.* **2**, eaam8638 (2017).

10. M. Jing, Beijing to release national artificial intelligence development plan. *South China Morning Post* (2017), (available at http://www.scmp.com/tech/article/2078209/beijing-release-national-artificial-intelligence-development-plan).