

11 August 2023

Level 11, 257 Collins Street
Melbourne VIC 3000
PO Box 38
Flinders Lane VIC 8009
T: (03) 8662 3300

Technology Strategy Branch
Department of Industry, Science and Resources
GPO Box 2013
Canberra ACT 2601

Submitted via upload to: <https://consult.industry.gov.au/supporting-responsible-ai>

Dear Technology Strategy Branch,

APS Response to the Safe and Responsible AI in Australia discussion paper

The Australian Psychological Society (APS) appreciates the opportunity to respond to the Commonwealth Department of Industry, Science and Resources' discussion paper about *Safe and Responsible AI in Australia*. Rather than addressing the questions in the discussion paper in detail, many of which are outside of the remit of the APS, we have provided broader insights and recommendations about safe and responsible AI.

The APS is the leading professional body for psychologists in Australia. We are dedicated to advancing the scientific discipline and ethical practice of psychology and work to realise the full human potential of individuals, organisations and their communities through the application of psychological science and knowledge. We are informed in our work by the United Nations' Sustainable Development Goals which champion inclusivity, social equity and the empowerment of marginalised and vulnerable groups¹. To this end, we advocate on behalf of our diverse profession and community for the meaningful design and reform of Australian health, education and other systems that impact our society.

AI is one such system that has been augmenting and impacting many aspects of our lives for decades. With the latest tranche of generative AI and the anticipated speed of further AI development, ethical and safety concerns are giving rise to society-wide questions about protections that are needed to ensure privacy, transparency and equity to engender trust in AI.

The APS believes that AI has the potential to reap considerable benefits for humanity, including improved health, wellbeing and human potential. Artificial intelligence is already having a significant impact on mental health care (e.g.^{2,3}). Safeguarding mechanisms must, however, keep up with AI advancements to keep individuals and society safe. Our position is that psychologists possess extensive expertise to contribute to this subject and must play a key role in the development of rapidly evolving AI technologies and shaping the guardrails that govern their use in Australia.

AI is profoundly intertwined with psychology

AI is psychologically and socially significant. AI draws inspiration from human intellectual capacities, gains its strength from human data about how people think, feel and act, and proves valuable solely through creating a positive impact on the human experience. Understanding human psychology and the impact of AI on human emotions, skills, relationships and well-being is the basis of human-AI interaction and is essential to designing and regulating AI systems that are user-friendly but also safe and which do no harm.

As the capacities of AI technologies are outstripping our understanding of how they work (e.g.,⁴), awareness of the potential for maladaptive outcomes and harm to individuals and societies has been heightened. It is well publicised, and described in the discussion paper, that AI algorithms can perpetuate existing biases present in the data they are trained on and widen disparities for already vulnerable groups⁵.

Psychology can offer evidence-based insights into understanding and addressing bias and fairness and other potential psychological harms associated with AI⁶, for example AI anxiety⁷ and psychological dependency on AI chat bots⁸ (and see Box 1).

Thus, in addition to considering the technical aspects of AI, policymakers and regulators must be aware of the extent of social and psychological benefits and harms associated with AI as they design, implement and evaluate AI safeguards. Integrating insights from psychological science and bringing together psychological practitioners and researchers, computer scientists, other professional stakeholders and consumers to work closely with regulators and policymakers over time will be essential for a safe and trusted AI ecosystem that aligns with human values and societal needs.

Box 1: Examples of Psychological Contributions for Trustworthy and Safe AI Systems

With more than a century of research, practice and insight into the human mind and the prediction of human behaviour, psychology is well positioned to make significant contributions to the development, deployment and oversight of trustworthy and safe AI systems.

Decision Making: Use of AI is not new in the mental setting - psychology has been contending with algorithmic versus clinical (human) decision making since the 1950s⁹. Psychology has also contributed a deep understanding of ethics and morality in diverse contexts, including but not limited to psychological practice and research (see ¹⁰) and is well placed to inform the development and regulation of AI systems that align with human values and ethical principles (e.g. ¹¹).

Bias and Fairness: Psychology has a long history of studying cognitive biases and decision-making heuristics and how these manifest and affect different groups of people (see ^{12,13}). Psychologists can support AI developers to address biases in training data and algorithms and co-design with regulators guidelines, standards and other mechanisms such as psychological audits to ensure fairness, transparency, and accountability in the development, deployment and oversight of generative AI systems^{14,15}.

Explainability: AI systems often operate as black boxes, making it challenging for users to understand their decision-making processes. Psychology can also help design AI systems that enable transparency with regard to decision-making including error types and error rates and psychological mechanisms associated with trust violation and trust repair¹⁶⁻¹⁸.

Human Displacement: AI technologies create concern about human displacement and redundancy, for example changed roles, job loss and the impact on the workforce landscape, as demonstrated in recent Australian data showing that less than one in four people believe AI and automation will create more jobs than are eliminated¹⁹. Psychologists' expertise in areas such as the psychology of work and organisations can support AI developers and regulators in designing, deploying and overseeing systems that recognise and respond to address concerns and uphold human rights to decent and meaningful work in the context of AI developments²⁰⁻²³.

Considerations for AI regulation

The design of regulatory approaches must also be informed by our understanding of psychology and human behaviour. For example, it may be tempting to rely on warnings, notices and other information-based regulatory devices (especially under self-regulatory approaches). While these techniques are relatively simple to create and enforce, the psychological effects must be considered.

Information (including warnings or notices) must be likely to be understood by the person in the context in which it is provided. Information provided under heavy cognitive load (e.g., where the person is already processing other complex or novel information) is unlikely to be perceived and retained in a meaningful way. Similarly, the way in which information is presented must consider how people differentially respond to the framing and communication of risk, probability and uncertainty. It would be undesirable for regulatory devices to transfer the burden of risk back to the individual through neglecting psychological principles.

In addition, the APS is supportive of a system of AI safeguarding that includes a number of elements such as:

- A risk-based approach for AI regulation that:
 - focuses on identifiable harms – including longer-term and subtler psychological effects of AI use,
 - holds AI developers accountable/liable in cases of misuse, and
 - includes psychologists as an integral part of any impact assessment process (as proposed in the discussion paper).

We note that risk-based regulatory approaches are likely to be less effective at the level of more subtle psychological effects of AI (e.g., the role of AI and algorithms in changing people's preferences over time and potential for harms such as radicalisation).

- Industry guidelines, standards and frameworks (for example, health, mental health, human resources) directly informed by industry peaks and professionals that can help to guide the safe use of available AI systems in a way that minimises the potential for negative psychological and social impacts.

While the APS supports the development of industry-developed guidelines, standards, and frameworks for AI use as part of the safeguarding system, we do not support transferring the burden/risk for adherence to AI regulation to industry groups and professionals. This would be unfairly burdensome for non-developers of AI with regard to time and resources required to keep up with a rapidly evolving field.

- Educating the public about the potential benefits and harms of AI, personal steps they can take to mitigate them and the role of the safeguarding system to protect society and prevent harm to individuals and groups.

Educating people about AI risks requires a deep understanding of the psychology of how people respond differentially to risk information. Informational responses such as warnings or notices need to consider that people will not necessarily respond in a rational way and that the salience of this messaging will diminish over time.

Policymakers must ensure that regulatory touchpoints such as those above are developmentally, cognitively, and culturally appropriate. Complex technical information and principles relating to AI transparency, explainability and training must be tailored in such a way as to make the nature of the safeguards accessible to a wide range of users, and particularly for vulnerable individuals and communities.

Finally, we take this opportunity to emphasise that while we embrace validated AI advancements in health and mental health care, we remain wary of the use of AI in ways that overlook essential human needs and care system shortcomings.

The APS would welcome the opportunity to bring our extensive expertise to the table and collaborate with regulators and industry on developing the Australian regulatory AI system, including industry-related guidelines, standards and frameworks.

If any further information is required from the APS, I would be happy to be contacted through the national office on (03) 8662 3300 or by email at z.burgess@psychology.org.au

Yours sincerely,

Dr Zena Burgess, FAPS FAICD
Chief Executive Officer

References

1. United Nations Department of Economic and Social Affairs. (2022). *Sustainable development*. <https://sdgs.un.org/>
2. *Wysa*. (n.d.). <https://www.wysa.com/>
3. Boucher, E. M., Harake, N. R., Ward, H. E., Stoeckl, S. E., Vargas, J., Minkel, J., Parks, A. C., & Zilca, R. (2021). Artificially intelligent chatbots in digital mental health interventions: A review. *Expert Review of Medical Devices*, *18*(sup1), 37–49. <https://doi.org/10.1080/17434440.2021.2013200>
4. Naughtin, C., & Bentley, S. V. (2023, June 16). Both humans and AI hallucinate—But not in the same way. *The Conversation*. <http://theconversation.com/both-humans-and-ai-hallucinate-but-not-in-the-same-way-205754>
5. Thomas, J., McCosker, A., Parkinson, S., Hegarty, K., Featherstone, D., Kennedy, J., Holcombe-James, I., Ormond-Parker, L., & Ganley, L. (2023). *Measuring Australia's digital divide: The Australian digital inclusion index 2023*. ARC Centre of Excellence for Automated Decision-Making and Society, RMIT University, Swinburne University of Technology, and Telstra. <https://doi.org/10.25916/528S-NY91>
6. Abrams, Z. (2023, August). AI is here. *Monitor on Psychology*, 46–53.
7. Li, J., & Huang, J.-S. (2020). Dimensions of artificial intelligence anxiety based on the integrated fear acquisition theory. *Technology in Society*, *63*, 101410.
8. Xie, T., Pentina, I., & Hancock, T. (2023). Friend, mentor, lover: Does chatbot engagement lead to psychological dependence? *Journal of Service Management*, *34*(4), 806–828. <https://doi.org/10.1108/JOSM-02-2022-0072>
9. Westen, D., & Weinberger, J. (2005). In praise of clinical judgment: Meehl's forgotten legacy. *Journal of Clinical Psychology*, *61*(10), 1257–1276. <https://doi.org/10.1002/jclp.20181>
10. Ellemers, N., Pagliaro, S., & Nunspeet, F. van (Eds.). (2024). *The Routledge international handbook of the psychology of morality*. Routledge.
11. Hagendorff, T. (2023). AI ethics and its pitfalls: Not living up to its own standards? *AI and Ethics*, *3*(1), 329–336. <https://doi.org/10.1007/s43681-022-00173-5>
12. Gilovich, T., Griffin, D., & Kahneman, D. (2002). *Heuristics and biases: The psychology of intuitive judgment*. Cambridge University Press.
13. Sibley, C. G., & Barlow, F. K. (Eds.). (2016). *The Cambridge Handbook of the Psychology of Prejudice* (1st ed.). Cambridge University Press. <https://doi.org/10.1017/CBO9781316161579>
14. Landers, R. N., & Behrend, T. S. (2023). Auditing the AI auditors: A framework for evaluating fairness and bias in high stakes AI predictive models. *American Psychologist*, *78*(1), 36–49. <https://doi.org/10.1037/amp0000972>
15. Bigman, Y. E., Wilson, D., Arnestad, M. N., Waytz, A., & Gray, K. (2023). Algorithmic discrimination causes less moral outrage than human discrimination. *Journal of Experimental Psychology: General*, *152*(1), 4–27. <https://doi.org/10.1037/xge0001250>
16. Langer, M., König, C. J., Back, C., & Hemsing, V. (2023). Trust in Artificial Intelligence: Comparing trust processes between human and automated trustees in light of unfair bias. *Journal of Business and Psychology*, *38*(3), 493–508. <https://doi.org/10.1007/s10869-022-09829-9>
17. Lacroux, A., & Martin-Lacroux, C. (2022). Should I trust the Artificial Intelligence to recruit? Recruiters' perceptions and behavior when faced with algorithm-based recommendation systems during resume screening. *Frontiers in Psychology*, *13*, 895997. <https://doi.org/10.3389/fpsyg.2022.895997>
18. Binz, M., & Schulz, E. (2023). Using cognitive psychology to understand GPT-3. *Proceedings of the National Academy of Sciences*, *120*(6), e2218523120.

19. Biddle, N. (2023). *Views of Australians towards science and AI* (Australia) [Report]. Centre for Social Research and Methods. <https://apo.org.au/node/323444>
20. Alcover, C.-M., Guglielmi, D., Depolo, M., & Mazzetti, G. (2021). Aging-and-Tech Job Vulnerability: A proposed framework on the dual impact of aging and AI, robotics, and automation among older workers. *Organizational Psychology Review, 11*(2), 175–201. <https://doi.org/10.1177/2041386621992105>
21. Blustein, D. L., Kenny, M. E., Di Fabio, A., & Guichard, J. (2019). Expanding the impact of the psychology of working: Engaging psychology in the struggle for decent work and human rights. *Journal of Career Assessment, 27*(1), 3–28. <https://doi.org/10.1177/1069072718774002>
22. Tang, P. M., Koopman, J., Mai, K. M., De Cremer, D., Zhang, J. H., Reynders, P., Ng, C. T. S., & Chen, I.-H. (2023). No person is an island: Unpacking the work and after-work consequences of interacting with artificial intelligence. *Journal of Applied Psychology*. <https://doi.org/10.1037/apl0001103>
23. Bankins, S., & Formosa, P. (2020). When AI meets PC: Exploring the implications of workplace social robots and a human-robot psychological contract. *European Journal of Work and Organizational Psychology, 29*(2), 215–229. <https://doi.org/10.1080/1359432X.2019.1620328>