Moral Responsibility in Complex Hybrid Intelligence Systems

David Lyreskog[1,2][0000-0001-6888-6272] , Hazem Zohny,[2,3][0000-0002-7734-2186], [1,3]Edmond Awad [1,3][0000-0001-7272-7186], Julian Savulescu [2,3,4][0000-0003-1691-6403], Ilina Singh [1,2][0000-0003-4497-3587]

[1] NEUROSEC, Department of Psychiatry, University of Oxford, Oxford, OX3 7JX, UK
[2] Wellcome Centre for Ethics & Humanities, Oxford, OX3 7LF, UK
[3] Oxford Uehiro Centre for Practical Ethics, Oxford, OX1 1PT, UK
[4]National University of Singapore, #02-03, 10 Medical Drive, 117597, Singapore
david.lyreskog@psych.ox.ac.uk

Abstract. In this paper, we describe how Hybrid Intelligence (HI) networks can be understood and analysed as Complex Systems, and show how traditional theories and methods of ascribing (moral) responsibility in the field of ethics and technology provide inadequate guidance with regard to responsibility distribution in Complex HI (CHI) Systems. Building on recent work in this area (1) we argue this is primarily due to the tendency of those theories to insist on either (A) individual-level responsibility distribution, or (B) collective-level responsibility distribution frameworks, relying on clear distinctions between individuals and collectives, as well as the presence of joint intentions. CHI Systems, we argue, do not easily lend themselves to be described in this way, and therefore our understanding of responsibility distribution ought to be adapted.

We propose a path away from traditional methods of responsibility distribution, to facilitate ethical analysis of HI networks. In doing so, we first explore ways to achieve moral responsibility analysis of CHI Systems, using relational accounts of autonomy and identity as basis. We then propose a framework of relational responsibility to be applied, with an emphasis on forward-looking responsibility and dynamic improvement. We conclude by discussing the strengths and weaknesses of such an approach, and stake out a way forward for ethical Complex Hybrid Intelligent Systems.

Keywords: Moral Responsibility, Ethics, Hybrid Intelligence, Complex Systems

1 Introduction

In this paper, we describe how Hybrid Intelligence (HI) networks can be understood and analysed as Complex Systems, and show how traditional theories and methods of ascribing (moral) responsibility in the field of ethics and technology provide inadequate guidance with regard to responsibility distribution in Complex HI (CHI) Systems. Building on recent work in this area (1) we argue this is primarily due to the tendency of those theories to insist on either (A) individual-level responsibility distribution, or (B) collective-level responsibility distribution frameworks, relying on clear distinctions between individuals and collectives, as well as the presence of joint intentions. CHI Systems, we argue, do not easily lend themselves to be described in this way, and therefore our understanding of responsibility distribution ought to be adapted.

We propose a path away from traditional methods of responsibility distribution, to facilitate ethical analysis of HI networks. In doing so, we first explore ways to achieve moral responsibility analysis of CHI Systems, using relational accounts of autonomy and identity as basis. We then propose a framework of relational responsibility to be applied, with an emphasis on forward-looking responsibility and dynamic improvement. We conclude by discussing the strengths and weaknesses of such an approach, and stake out a way forward for ethical Complex Hybrid Intelligent Systems.

2 Background

2.1 Hybrid Intelligence

Hybrid Intelligence (HI) typically denominates the combination of human intelligence and Artificial Intelligence (AI) to achieve some more or less defined goal.[1] HI networks involve the integration of human cognitive abilities – such as intuition, creativity, and critical thinking – with machine learning algorithms, natural language processing, and other AI technologies. Examples of HI include human-assisted AI, where humans work alongside machines to train AI algorithms, or, conversely, AI-assisted human decision-making, where AI provides data driven insights and recommendations to assist decision -making [2,3,4,5].

HI systems can be as small or large as they need to be, and may contain as many human or AI components as seen fit, with information flowing in few or many directions. In other words, they can have low or high *directionality*. Furthermore, HI systems can be more or less *direct*. For instance, while a human using an online chatbot to help write an article abstract is an example of indirect collaboration, a brain-computer interface with feedback loops between a language model and a human brain may be considered more direct. In many cases, it seems that the more direct and highly directional a system becomes, the more difficult it is to track and dissect the intentions and contributions of each individual component – in particular where the network has synergetic effects, leading to outcomes which do not easily translate to the sum of the contributions of the constituents(1). Such an HI network can be analysed as a Complex System.

[1] Ambient- and Augmented Intelligence, or other forms of intelligence, could arguably also be part of HI constellations. However, for the sake of clarity, we here focus on AI and human intelligence as key components.

2.2 Complex Systems

A Complex System can be understood as a collection or compound of interacting components or agents exhibiting complex behaviour. Complex Systems are typically characterized by their intricate dynamics, where small changes in one component or agent can have significant effects on the entire system. Examples of Complex Systems can be found in many different fields of study, including biology, ecology, economics, and physics, but can also be found in other domains, such as, financial markets, social networks, or indeed the human brain.[6,7,8,9,10,11,12].

Crucially for our purposes here, Complex Systems have emergent properties: the behaviour of the system as a whole cannot be exhaustively described or understood solely by observing the behaviour of its individual agents or constituents in isolation. Instead, the properties of the system arise from the interactions and feedback loops between the constituents. Furthermore, Complex Systems are adaptive, capable of changing their behaviour in response to changing environments and inputs. This adaptability make Complex Systems resilient in the face of change or disruption, but can also make their mechanisms and functions elusive and difficult to analyse. Therefore, where HI networks are aptly understood as Complex Systems, analysing distribution of responsibility within that network (or between networks) can prove challenging.

2.3 Moral Responsibility and CHI Systems

Most commonly, responsibility is attributed on an individual level: e.g., *a person P is responsible for an outcome O if P caused O, P knew what O would entail, and P reasonably could have acted otherwise (not O)*. In other cases, P cannot, and/or does not, act alone, but does so together with others. In such cases, we may be inclined to ascribe joint, or collective responsibility. Most ethical frameworks for collective responsibility require that (i) all members of the collective are aware that they are part of that collective, and identify as members of that collective; (ii) all members intend to cause an outcome as that specific collective, and (iii) all members contribute to the realization of that outcome [1,13,14,15]. However, we argue, these criteria may not be met in CHI Systems, and yet we may hold the alternatives – individual responsibility, or no responsibility at all – to be unsavoury and/or counterintuitive. For instance, if a large CHI System causes a catastrophic outcome due to the effects of an emergent property, it will be difficult to fairly pinpoint which individual(s) to blame. Yet, it seems CHI Systems and their constituent agents are not necessarily amoral. Moral responsibility attribution in CHI Systems may instead be better understood and served by an approach which focuses on the relational aspects of responsibility – an approach which is increasingly used in ethics analysis and assessment in health and care settings to understand complex multi-agent decision-making and action [16,17.18].

3 Methods

In this paper we adopt a conceptual analysis methodology [19,20] to the concept(s) of '(individual, collective, relational) responsibility', and the related terminologies and taxonomies. We then apply a normative analysis methodology [21,22] to show how a coherent framework for relational responsibility distribution and attribution can inform CHI Systems development, policy, and praxis.

4 Results

4.1 Conceptual analysis

We focus on a select number of features of moral responsibility, which we take to be key to understanding CHI Systems. First, it is important to distinguish *moral* responsibility from *causal* responsibility, and *legal* responsibility. While these are often connected (i.e. if an agent causes a bad outcome by acting immorally, they may face legal consequences), they are distinct concepts which require separate attention. In this paper, we focus our attention specifically on moral responsibility, while acknowledging the role of causal and legal responsibility in shaping a practical ethic for CHI Systems.

Second, we distinguish between backward-looking (moral) responsibility, and forward-looking responsibility. Backward-looking responsibility typically ascribes and/or describes attribution of *blame* and *praise* for something which has occurred, while forward-looking responsibility ascribes and/or describes attribution of *duty* and *incentive* for some state of affairs which is either to be maintained, or (perhaps more commonly) to be brought about [23].

Thirdly, we outline the key features of the concept 'relational responsibility'[24], and its sister concepts 'relational identity'[25] and 'relational autonomy' [26,27,28]. We emphasize the necessity of relationality as a general component of moral responsibility. E.g., '*Person P is backward-looking responsible for action*[2] *X to Person Q means that it is fitting for Q to hold P responsible for X'*, and *'P is forward-looking responsible for X to Q means that P owes it to Q to see to it that X'*. In the context of CHI Systems, we further note that *P* can co-constitute *Q*, and vice versa, as well as the possibility that *P=Q* under some circumstances.

4.2 Normative analysis

(Moral) responsibility ascription and attribution in AI and HI networks being difficult (not to say impossible) is widely considered to be a problem [29,30,31]. This problem is largely grounded in applied responsibility theory being inapt in the domain in question, as it fails to speak to our intuitions and experiences about actions and omissions in HI networks. To address this problem, there is a case to be made to analyse HI networks as Complex (CHI) Systems. To analyse CHI Systems, one needs to take into account not only the agents and their individual or collective attributes, but also the dynamic interrelations between them. To inform the development of a framework for moral responsibility ascribing and distributing in CHI Systems, we can take inspiration and guidance from other domains and sectors. In particular, there is a growing bioethics literature on the role and importance of relationships in health and

[2] 'X' could denote an outcome rather than an action, depending on which framework we apply.

care, spanning relational identity [25, 32,33], relational autonomy [26, 27, 28, 34, 35], and relational responsibility [24, 36, 37]. Relational identity can be understood as aspects of identity retention and shaping which to some degree are determined by persons' relationships with others (e.g. family/spousal care of persons living with dementia); relational autonomy typically denotes interrelational aspects of self-determination and decision-making (e.g. shared decision-making in complex medical dilemmas); relational responsibility, subsequently, can be understood as a framework for ascribing and distributing responsibility within multi-agent networks as a "means of valuing, sustaining, and creating forms of relationship out of which common meanings – and thus moralities – can take wing" [24].

In taking a relational approach to analysing responsibility in CHI Systems, notably, forward-looking responsibility takes a dominant role, at the expense of backward-looking responsibility. This is due to at least two factors. First, relational responsibility frameworks emphasise relationship building and cultivation. While backward-looking responsibilization mechanisms (such as blaming or praising) play a part in such cultivation processes, they appear neither sufficient nor necessary for such a project. Forward-looking responsibility, on the other hand, appears to play a central role in relationship cultivation. Second, analysis of networks and systems is normally focused on improvement. Therefore, an analysis of moral responsibility in HI networks as Complex Systems naturally lends itself to a more forward-looking structure aimed at improvement of those systems.

5 Conclusion

HI networks are often constituted by a large number of agents, some of which we would identify as moral agents (competent persons), and some which we would ascribe an amoral or non-moral status (XI technology). Widely accepted frameworks for ascribing and distributing moral responsibility are poorly equipped to aptly analyse these HI networks. One reason for this is that HI systems need to be understood as Complex Systems, and most ethical frameworks struggle with such systems. We propose that HI networks should be understood and analysed as Complex Systems, and that relational responsibility accounts of moral responsibility are better equipped than other frameworks to perform ethical analysis and guide development, policy, and praxis in this domain. This will have profound impacts on how we understand ethical behaviour in HI, in particular as moral responsibility may give more weight to forward-looking responsibility than to backward-looking responsibility.

References

1. Lyreskog DM, Zohny H, Savulescu J, Singh I. Merging Minds: The Conceptual and Ethical Impacts of Emerging Technologies for Collective Minds. Neuroethics. 2023 Apr;16(1):12.
2. Tvoroshenko I, Gorokhovatskyi V. The Application of Hybrid Intelligence Systems for Dynamic Data Analysis.Author, F., Author, S., Author, T.: Book title. 2nd edn. Publisher, Location (1999).
3. Sayin B, Krivosheev E, Ramírez J, Casati F, Taran E, Malanina V, Yang J. Crowd-Powered Hybrid Classification Services: Calibration is all you need. In2021 IEEE International Conference on Web Services (ICWS) 2021 Sep 5 (pp. 42-50). IEEE.
4. Ye P, Wang X, Zheng W, Wei Q, Wang FY. Parallel cognition: Hybrid intelligence for human-machine interaction and management. Frontiers of Information Technology & Electronic Engineering. 2022 Dec;23(12):1765-79.
5. Wiethof C, Bittner EA. Toward a hybrid intelligence system in customer service: collaborative learning of human and AI.
6. Ma'ayan, A., 2017. Complex systems biology. *Journal of the Royal Society Interface*, *14*(134), p.20170391.
7. Jørgensen SE, Patten BC, Straškraba M. Ecosystems emerging: toward an ecology of complex systems in a complex future. Ecological Modelling. 1992 Jul 1;62(1-3):1-27.
8. Gomes O, Gubareva M. Complex systems in economics and where to find them. Journal of Systems Science and Complexity. 2021 Feb;34(1):314-38.
9. Holovatch Y, Kenna R, Thurner S. Complex systems: physics beyond physics. European Journal of Physics. 2017 Feb 15;38(2):023002.
10. Kuhlmann M. Explaining financial markets in terms of complex systems. Philosophy of Science. 2014 Dec;81(5):1117-30.
11. Bento F, Tagliabue M, Sandaker I. Complex systems and social behavior: Bridging social networks and behavior analysis. Behavior science perspectives on culture and community. 2020:67-91.
12. Telesford QK, Simpson SL, Burdette JH, Hayasaka S, Laurienti PJ. The brain as a complex system: using network science as a tool for understanding the brain. Brain connectivity. 2011 Oct 1;1(4):295-308.
13. Beier K, Jordan I, Wiesemann C, Schicktanz S. Understanding collective agency in bioethics. Medicine, Health Care and Philosophy. 2016 Sep;19:411-22.
14. Tuomela R. We-intentions revisited. Philosophical Studies. 2005 Sep;125:327-69.
15. Gilbert M. The structure of the social atom: Joint commitment as the foundation of human social behavior. Socializing metaphysics. 2003:39-64.
16. Jacobs G. Patient autonomy in home care: Nurses' relational practices of responsibility. Nursing ethics. 2019 Sep;26(6):1638-53.
17. Tirri K, Husu J. Care and responsibility in'the best interest of the child': Relational voices of ethical dilemmas in teaching. Teachers and Teaching. 2002 Feb 1;8(1):65-80.
18. Allwood J, Johansson IL, Olsson LE, Tuna G. On the need for an ethical understanding of health-care accountability. Journal of Organisational Transformation & Social Change. 2015 Aug 1;12(2):121-37.
19. Jackson F. From metaphysics to ethics: A defence of conceptual analysis. Clarendon Press; 1998 Jan 8.
20. Olsthoorn J. Conceptual analysis. Methods in analytical political theory. 2017 May 2:153-91.
21. Daniels N. Justice and justification: Reflective equilibrium in theory and practice. Cambridge University Press; 1996 Sep 28.

22. Earp BD, Demaree-Cotton J, Dunn M, Dranseika V, Everett JA, Feltz A, Geller G, Hannikainen IR, Jansen LA, Knobe J, Kolak J. Experimental philosophical bioethics. AJOB Empirical Bioethics. 2020 Jan 2;11(1):30-3.
23. Vincent NA, Van de Poel I, Van Den Hoven J, editors. Moral responsibility: Beyond free will and determinism. Springer Science & Business Media; 2011 Aug 17.
24. McNamee S, Gergen KJ. Relational responsibility: Resources for sustainable dialogue. Sage; 1999.
25. Koster L, Ybema SB, Jonkers IR. I've got you under My skin: Relational identity work in interactional dynamics. InAcademy of Management Proceedings 2018 Jul 2 (Vol. 2018, No. 1, p. 14694). Briarcliff Manor, NY 10510: Academy of Management.
26. Mackenzie C, Stoljar N, editors. Relational autonomy: Feminist perspectives on autonomy, agency, and the social self. Oxford University Press; 2000 Jan 27.
27. Ho A. Relational autonomy or undue pressure? Family's role in medical decision-making. Scandinavian journal of caring sciences. 2008 Mar;22(1):128-35.
28. Christman J. Relational autonomy, liberal individualism, and the social constitution of selves. Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition. 2004 Jan 1;117(1/2):143-64.
29. Hakli R, Mäkelä P. Moral responsibility of robots and hybrid agents. The Monist. 2019 Apr 1;102(2):259-75.
30. Loh W, Loh J. Autonomy and responsibility in hybrid systems. Robot ethics. 2017 Sep 1;2.
31. Akata Z, Balliet D, De Rijke M, Dignum F, Dignum V, Eiben G, Fokkens A, Grossi D, Hindriks K, Hoos H, Hung H. A research agenda for hybrid intelligence: augmenting human intellect with collaborative, adaptive, responsible, and explainable artificial intelligence. Computer. 2020 Aug 1;53(08):18-28.
32. Lindemann, H., 2016. *Holding and letting go: The social practice of personal identities*. Oxford University Press.
33. Baylis, F., 2013. "I am who I am": On the perceived threats to personal identity from deep brain stimulation. *Neuroethics*, *6*, pp.513-526.
34. Goering, S., Klein, E., Dougherty, D.D. and Widge, A.S., 2017. Staying in the loop: Relational agency and identity in next-generation DBS for psychiatry. *AJOB Neuroscience*, *8*(2), pp.59-70.
35. Gómez-Vírseda, C., De Maeseneer, Y. and Gastmans, C., 2019. Relational autonomy: what does it mean and how is it used in end-of-life care? A systematic review of argument-based ethics literature. *BMC medical ethics*, *20*(1), pp.1-15.
36. Sherwin, S. and Stockdale, K., 2017. Whither bioethics now? The promise of relational theory. *IJFAB: International Journal of Feminist Approaches to Bioethics*, *10*(1), pp.7-29.
37. Gary, M., 2023. Relational approaches in bioethics: A guide to their differences. *Bioethics*.