BJHS Themes (2023), **8**, 189–203
doi:10.1017/bjt.2023.8

RESEARCH ARTICLE

The 'artificial intelligentsia' and its discontents: an exploration of 1970s attitudes to the 'social responsibility of the machine intelligence worker'

Rosamund Powell

The Alan Turing Institute, UK
Email: rpowell@turing.ac.uk

Abstract

In 1972, ten members of the machine intelligence research community travelled to Lake Como, Italy, for a conference on the 'social implications of machine intelligence research'. This paper explores their varied and contradictory approaches to this topic. Researchers, including John McCarthy, Donald Michie and Richard Gregory, raised 'ethical' questions surrounding their research and actively predicted risks of machine intelligence. At the same time, they delayed any action to mitigate these risks to an uncertain future where technical capabilities were greater. I argue that conference participants' claims that 1972 was 'too early' to speculate on societal impacts of their research were disingenuous, motivated both by threats to funding and by researchers' own politically informed speculation on the future.

In June of 1972, ten prominent men drawn from the machine intelligence research community and interconnected industries came together at Villa Serbelloni on the shores of Lake Como, Italy. Their purpose was to address threats to humanity which might arise due to machine intelligence, a term they used interchangeably with AI to define machines with human-comparable intelligence.[1] These men, henceforth the 'Serbelloni group', were Cordell Green (Stanford), Peter Landin (Queen Mary College), Donald Michie (Edinburgh University), John Alan Robinson (Syracuse University), Robert Taylor (Xerox Palo Alto Research Centre), Peter Will (IBM), John McCarthy (MIT), Daniel Bobrow (Xerox Palo Alto Research Centre), Takayasu Ito (Mitsubishi Electric Corporation) and Lord Balfour of Burleigh (Bank of Scotland).[2] I draw on previously unexplored archival materials detailing this conference to show that machine intelligence researchers actively pursued 'ethical questions' surrounding their research during the 1970s.[3]

[1] Attendees used 'machine intelligence' synonymously with 'AI'. I use the term 'machine intelligence' to reflect the language used at Villa Serbelloni itself while also referencing 'AI' to reflect on relevant historical treatments. Throughout, 'machine intelligence' and 'AI' are used as placeholders for multiple overlapping genealogies of technique. To deal with these genealogies would go beyond my scope, which focuses instead on the parameters of 'social implications'.

[2] Rockefeller Foundation, 'The social implications of machine intelligence research Villa Serbelloni, 22021 Bellagio (Como), Italy 11–15 June 1972' (manuscript), Rockefeller Foundation Collection, Machine Intelligence Conference, RF RG 1.3, Series 120, Box 40, Folder 243, available via Rockefeller Archive Center, New York.

[3] Richard Gregory, 'Social implications of intelligent machines', in Bernard Meltzer and Donald Michie (eds.), *Machine Intelligence 6*, Edinburgh: Edinburgh University Press, 1971, pp. 3–13, 12.

At times, to borrow a phrase from historian Shunryu Colin Garvey, they acted as 'discontents', defined throughout as those who 'doubt, question, challenge, reject, reform and otherwise reprise "AI"'.[4] Nevertheless, they made contributions that were in several ways characteristic of a more unified 'AI establishment', a term introduced by James Fleck to map a set of interconnected researchers who approach a sufficient 'degree of commonality to be called a community'.[5]

In this paper I uncover the significant role played by 'establishment' figures in initiating discourse on social responsibility, but also characterize their contributions and expose their collective desire to postpone the very discourse they had themselves actively set out to engage in. On the one hand, they questioned the future they would create through their research. For example, they cited concerns over how machine intelligence might be used by political megalomaniacs. On the other hand, the parameters of their debate were shaped by funding cycles and technological determinism, and, in the end, they advocated postponing action until an arbitrary future when AI capabilities had advanced. Their narrow focus on a future where machines had achieved human-comparable intelligence has striking parallels today, as many researchers continue to be preoccupied by the societal impact of future AGI (artificial general intelligence), neglecting more immediate harms which might occur along the way.

The role played by Donald Michie in organizing this conference is particularly significant and complicates existing accounts of Michie as a staunch defender of machine intelligence research. Ultimately, the Serbelloni group did not offer Michie all he wanted from discussions on the social implications of machine intelligence research. Consequently, he increasingly turned to historians and politicians for advice and took a path that was significantly influenced by his socialist politics.[6] He did so while simultaneously remaining a key figure in the social web of the AI establishment.

A historiography of pioneers and critics

Concern surrounding the societal implications of new technologies was by no means new in 1972, even for members of the AI research community. For example, Marvin Minsky and Herbert Simon contributed to *Computers and the World of the Future*, a 1964 volume exploring the impact of computing on society.[7] Nevertheless, to the extent that historical accounts have addressed 1970s critique on AI and society from within machine intelligence research communities, they have focused on a 'proud heretic' dissenting against his research community.[8] Namely it has been presented as the domain of Joseph Weizenbaum, whose 1976 book *Computer Power and Human Reason* condemned machine intelligence researchers for failing to question their impact on humanity and cast this loosely defined, elite group as the 'artificial intelligentsia', thus critiquing the social structure of the community to which he belonged as a computing professor at MIT.[9] The

[4] Shunryu Colin Garvey, 'Unsavory medicine for technological civilization: introducing "Artificial Intelligence & its Discontents"', *Interdisciplinary Science Reviews* (2021) 46(1–2), pp. 1–18, 2.

[5] James Fleck, 'Development and establishment in artificial intelligence', in Norbert Elias, Herminio Martins and Richard Whitley, *Scientific Establishments and Hierarchies*, London: D. Reidel, 1982, pp. 169–219, 169.

[6] Donald Michie, correspondence with Balfour on Spetsai conference, April 1975, Donald Michie Collection, MS 88958/1/458, available via British Library: London.

[7] Martin Greenberger (ed.), *Computers and the World of the Future*, Cambridge, MA: MIT Press, 1964.

[8] Zachary Loeb, 'The lamp and the lighthouse: Joseph Weizenbaum, contextualizing the critic', *Interdisciplinary Science Reviews* (2021) 46(1–2), pp. 19–35.

[9] Joseph Weizenbaum, *Computer Power and Human Reason: From Judgement to Calculation*, Harmondsworth: Penguin, 1976.

https://doi.org/10.1017/bjt.2023.8 Published online by Cambridge University Press

so-called 'artificial intelligentsia' then challenged Weizenbaum through reviews which defended their discipline.[10]

This dichotomy between a unified, defensive 'intelligentsia' and a 'lone heretic' emerging from within has been made possible by two separate historiographical trends. First, the earliest histories of AI, such as those written by Daniel Crevier and Pamela McCorduck, were written by insiders with a tendency to celebrate the hero pioneers of the discipline.[11] Such accounts lend themselves to a binary whereby the discipline of AI is celebrated, while Weizenbaum's critique is represented as an intrusion. Recent scholarship in the history of computing has taken issue with these internalist histories and proposed a revised approach whereby prominent researchers are critically contextualized.[12] The social implications of machine intelligence have been re-explored in this light as exploring critics of machine intelligence has become a research priority.[13] This second wave of scholarship has again shone a light on Weizenbaum's social critique. For example, historian Zachary Loeb emphasizes Weizenbaum's outsider influences and describes him as 'lonely' within his research community, as a computing professor himself.[14] Historical framings continue to generalize machine intelligence researchers as dismissive of the negative social implications of their work. The binary between critics and pioneers persists.

On the single occasion the Serbelloni conference is mentioned, in a historical treatment by Margaret Boden, the 'establishment' are once again cast as dismissive of any social implications that their work may have. Boden writes that while some researchers did meet on Lake Como, 'John McCarthy refused to join them', believing it too soon for speculation.[15] This is factually incorrect – McCarthy did attend as an expert speaker.[16] Additionally, this emphasis on researchers' indifference misconstrues the initiative required to institute the meeting. In short, the Serbelloni group complicates a persistent historiographical binary as their actions reveal that supposed *pioneers* of machine intelligence did at times act as *discontents*.

The road to Serbelloni

In the two years leading up to the Serbelloni conference, Donald Michie and the experimental psychologist Richard Gregory, his former colleague at Edinburgh University, began to explore ethical questions surrounding their work. The Serbelloni conference was proposed by Michie, a former Bletchley Park code breaker turned geneticist who by 1972 was the director of the Department of Machine Intelligence and Perception he had co-founded five years earlier with Gregory and the theoretical chemist Christopher Longuet-Higgins.[17]

[10] Benjamin Kuipers, John McCarthy and Joseph Weizenbaum, 'Computer power and human reason', *SIGART Newsletter* (1976) 58, pp. 4–13; Donald Michie, 'Computer power and human reason: from judgement to calculation: by Joseph Weizenbaum', *International Journal of Man-Machine Studies* (1976) 8(6), pp. 743–5.

[11] Pamela McCorduck, *Machines Who Think: A Personal Inquiry into the History and Prospects of Artificial Intelligence*, San Francisco: Freeman, 1979; Daniel Crevier, *AI: The Tumultuous History of the Search for Artificial Intelligence*, New York: BasicBooks, 1993.

[12] Paul N. Edwards, *The Closed World: Computers and the Politics of Discourse in Cold War America*, Cambridge, MA: MIT Press, 1996; Jonathan Penn, 'Inventing intelligence: on the history of complex information processing and artificial intelligence in the United States in the mid-twentieth century', PhD dissertation, University of Cambridge, 2020.

[13] Garvey, op. cit. (4), p. 2.

[14] Loeb, op. cit. (8), p. 32.

[15] Margaret Boden, *AI: Its Nature and Future*, Oxford: Oxford University Press, 2016, p. 164.

[16] Rockefeller Foundation, op. cit. (2).

[17] Donald Michie, *Donald Michie: On Machine Intelligence, Biology and More* (ed. Ashwin Srinivasan), Oxford: Oxford University Press, 2009.

https://doi.org/10.1017/bjt.2023.8 Published online by Cambridge University Press

To make the conference possible, Michie secured Rockefeller Foundation funding, following almost two years of correspondence from August 1970 to June 1972. This culminated in the foundation hosting the event at the Bellagio Centre, Villa Serbelloni.[18] During this period, the conference title changed from The Social Responsibility of the Machine Intelligence Worker to The Social Implications of Machine Intelligence Research, itself a significant decision which marks a shift in focus from questions of morality and responsibility towards consequences or implications. This new framing leant itself more to the speculative and predictive methods that were ultimately adopted at Villa Serbelloni and made room for the consideration of positive social implications, something which became crucial to the discussion at Villa Serbelloni.

Michie's role was to determine the scope of the conference and, as he saw it, to define the parameters of a new topic. However, by this time, there was already significant writing about the social implications of AI from within the technical community, in addition to fiction, the social sciences and the humanities.[19] Neither Michie, nor any other 'lone heretic', was raising this question anew, yet the papers Michie curated for the conference show a narrow appreciation of the existing debate as he focused largely on arguments circulating within his own community, machine intelligence researchers.

First, Michie considered Jack Good's 1966 'Speculations concerning the first ultraintelligent machine'. Social issues were acknowledged as Good described the 'possibility that the human race will become redundant' alongside 'other ethical problems'.[20] However, the paper's primary focus was not on 'ethical problems' but rather on research theories which may facilitate future ultraintelligence. Good was a close colleague and 'best friend' to Michie during their time at Bletchley Park and an integral member of the UK research community.[21] The inclusion of this paper suggests that Michie prioritized research from friends and colleagues, even in cases where these touched only briefly on 'social implications'. This is reinforced by Michie's inclusion of McCarthy and Hayes's 'Some philosophical problems from the standpoint of artificial intelligence' and Green's response to that paper, which detailed a particular method for constructing a 'question-answering system'.[22] Neither paper refers to social implications.

Michie also included Weizenbaum's 1972 paper 'On the impact of the computer on society'.[23] Weizenbaum bemoaned what he described as 'the typical essay' on computers which recommended relying on the 'computer scientist' themself as protector against potential harms.[24] He argued it was necessary to look to side effects of computing, and to the impact computers would have on 'man's image of himself'. Arguing against techno-solutionism he advocated instead for 'human answers' to human questions.[25] Michie's inclusion of Weizenbaum's paper does not indicate agreement – many Serbelloni

[18] John Marshall, Correspondence detailing a call with Michie on the Serbelloni conference, 20 August 1970, Rockefeller Foundation Collection, Machine Intelligence Conference, RF RG 1.3, Series 120, Box 40, Folder 243, available via Rockefeller Archive Center, New York.

[19] Michael Falk, 'Artificial stupidity', *Interdisciplinary Science Reviews* (2021) 46(1–2), p. 37.

[20] Jack I. Good, 'Speculations concerning the first ultraintelligent machine,' *Advances in Computers* (1966) 6(C), pp. 31–88.

[21] Jack I. Good, foreword, in Michie, op. cit. (17), pp. xi–xxix.

[22] John McCarthy and Patrick J. Hayes, 'Some philosophical problems from the standpoint of artificial intelligence', in Bernard Meltzer and Donald Michie (eds.), *Machine Intelligence 4*, Edinburgh: Edinburgh University Press, 1969, pp. 473–502; Cordell C. Green, 'Theorem-proving by resolution as a basis for question-answering systems', in Meltzer and Michie, op. cit., pp. 183–205.

[23] Donald Michie, funding proposal on social implications of research in machine intelligence, 1973, Donald Michie Collection, MS 88958/3/107, available via British Library, London.

[24] Joseph Weizenbaum, 'On the impact of the computer on society', *Science* (1972) 176(4035), pp. 609–14.

[25] Weizenbaum, op. cit. (24).

https://doi.org/10.1017/bjt.2023.8 Published online by Cambridge University Press

participants were to become his outspoken critics in 1976 – but it does demonstrate a willingness to consider more pessimistic projections.

However, beyond Weizenbaum there were already numerous analyses on the social foundations and implications of new technologies on which Michie could have drawn. For example, in 1970 Lewis Mumford published the second volume of *The Myth of the Machine*: *The Pentagon of Power*, detailing which levers of power within military–industrial states were influencing the development of technology.[26] Similarly, the literature on scientists' and technologists' social responsibility was much broader than Michie's selection would suggest, from Bertrand Russell's 1960 paper on 'The social responsibilities of scientists' to Norbert Wiener's 1960 paper 'Some moral and technical consequences of automation', and this context is crucial to demonstrating just how narrow the discussions of the Serbelloni group were.[27]

While Michie undertook this organizational and agenda-setting role, he had not written anything substantial on the social implications of machine intelligence and instead relied on ideas Richard Gregory had presented in his 1971 paper 'Social implications of intelligent machines'.[28] Gregory did not travel to Lake Como, perhaps because he had moved on from Edinburgh and from machine intelligence research by this stage. His role was nevertheless significant. His paper was not only presented to the group but also sent by Michie to the Rockefeller Foundation and selected experts as justification for conference funding.[29] Jacob Bronowski, then at the Salk Institute and consulted by the Rockefeller Foundation, wrote that if the conference was 'on the same theme and standard' it certainly merited support.[30]

Gregory brought together three angles from which he argued societal consequences of machine intelligence must be considered: psychology, engineering and moral sciences. First, he considered psychological definitions of intelligence, arguing that there was no agreed definition of intelligence and social consequences would vary dramatically depending on the nature of intelligence in machine form.[31] Gregory's engineering interests were also reflected as he linked social implications to previous fears surrounding the clock, steam engine, horse-drawn bus and petrol engine. He argued that 'important effects arise from completely unnoticed origins' and described the example of horse-drawn buses which led to poor housing being built in difficult-to-drain valleys, resulting in illness.[32] Consequently, the 'social history' of engineering could help identify how societal consequences arise unexpectedly.

In addition, Gregory had studied moral sciences at Cambridge. He later described being taught by Bertrand Russell at a time when Russell had become 'a bit bored with symbolic logic' and increasingly interested in 'ethics'.[33] Gregory emphasized that 'perhaps the most

[26] Lewis Mumford, *The Myth of the Machine: The Pentagon of Power*, New York: Harcourt, Brace, Jovanovich, 1970.

[27] Bertrand Russell, 'The social responsibilities of scientists: a scientist can no longer shirk responsibility for the use society makes of his discoveries', *Science* (1960) 131(3398), pp. 391–2.

[28] Gregory, op. cit. (3), pp. 3–13.

[29] Warren Weaver, correspondence with Jane Allen on funding the social implications conference, 16 June 1971, Rockefeller Foundation Collection, Machine Intelligence Conference, RF RG 1.3, Series 120, Box 40, Folder 243, available via Rockefeller Archive Center, New York; Robert Morrison, correspondence with Jane Allen on funding the social implications conference, 1 July 1971, Rockefeller Foundation Collection, Machine Intelligence Conference, RF RG 1.3, Series 120, Box 40, Folder 243, available via Rockefeller Archive Center, New York.

[30] Jacob Bronowski, correspondence with Miss Jane Allen on social implications conference, 26 June 1971, Rockefeller Foundation Collection, Machine Intelligence Conference, RF RG 1.3, Series 120, Box 40, Folder 243, available via Rockefeller Archive Center, New York.

[31] Gregory, op. cit. (3), p. 5.

[32] Gregory, op. cit. (3), pp. 5–8.

[33] Richard Gregory, 'An interview with Richard Gregory', *Cogito* (1991) 5(3) p. 123.

https://doi.org/10.1017/bjt.2023.8 Published online by Cambridge University Press

important questions here concern the ethics of responsibility'.[34] In a detailed example surrounding the prospect of a machine judge, Gregory clearly asserted that social implications cannot be resolved by quantifying outcomes. In his view, they were unquantifiable because they were moral.[35]

In contrast to Michie's insider status, Gregory's article was shaped by his situation on the periphery of machine intelligence research. In 1966 he had moved to Edinburgh from lecturing in the Department of Experimental Psychology at the University of Cambridge and by 1970 had moved once more, to become a professor of neuropsychology at Bristol.[36] Three years at Edinburgh had given him an insider view of machine intelligence, but his methods remained those of a relative outsider, continuing subsequently with psychological research on the human eye.[37] He later described his move to Edinburgh as 'naïve', noting that he had joined a research community he 'barely knew' studying a topic which he did not yet see as advanced.[38] His status on the periphery, one step into the community but retaining distinct opinions, broadened the scope of this paper as he pulled together these distinct threads from psychology, engineering, social history and moral sciences.

Despite his use of Gregory's paper as indicative of the conference themes, Michie's final plans for the conference participant list see him once again hint at diverse viewpoints only to invite friends and fellow establishment researchers.[39] Michie's attempts to obtain multidisciplinary contributors were limited. Michie emphasized in correspondence that his wife, Jean Hayes Michie, was attending as a 'bona fide member' of the conference.[40] His perseverance led William Olson, director of Villa Serbelloni, to conclude that there was 'no choice but to accept the legitimacy of this'.[41] This could indicate a multidisciplinary approach as Jean Hayes Michie was a psychologist at Strathclyde University. In later years she wrote with her husband on topics including 'human-centred design' in machine intelligence.[42] However, she was given a place neither on the panel nor as an expert speaker.[43] She was grouped with other participants' wives as Olson described all of them as 'non-conferees'.[44] This resulted in her being entirely erased from the outline of the conference written by Cordell Green, and consequently any contributions she made cannot easily be distinguished.

Only one member of Michie's panel lacked a machine intelligence background. Lord Balfour of Burleigh was invited specifically to provide a non-expert contribution, revealing Michie's willingness to look beyond the 'AI establishment'. Yet Balfour was not selected for leading another relevant field of study – at this time he was director of the Bank of Scotland – but rather for a 'personal interest' in machine intelligence.

[34] Gregory, op. cit. (3), pp. 9–12.

[35] Gregory, op. cit. (3), pp. 9–10.

[36] Fleck, op. cit. (5), pp. 169–219.

[37] Richard Gregory, *Eye and Brain: The Psychology of Seeing*, 3rd edn, London: Weidenfeld and Nicolson, 1997.

[38] Richard Gregory, 'Professor Richard Gregory Edinburgh 1967–1970: the birth of artificial intelligence', Media Central UCL (27 May 2008), at https://mediacentral.ucl.ac.uk/Play/62556 (accessed 22 March 2022).

[39] Michie, op. cit. (6).

[40] W. Olson, correspondence with Jane Allen Wives at Villa Serbelloni (manuscript), Rockefeller Foundation Collection, Machine Intelligence Conference, RF RG 1.3, Series 120, Box 40, Folder 243, available via Rockefeller Archive Center, New York, 10 February 1972.

[41] Olson, op. cit. (40).

[42] J.H. Michie and D. Michie, 'Simulator-mediated acquisition of a dynamic control skill', *AI & Society* (1998) 12 (1), pp. 71–7.

[43] Rockefeller Foundation, op. cit. (2).

[44] William Olson, correspondence with Ralph Richardson on the success of social implications conference, 15 June 1972, Rockefeller Foundation Collection, Machine Intelligence Conference, RF RG 1.3, Series 120, Box 40, Folder 243, available via Rockefeller Archive Center, New York.

https://doi.org/10.1017/bjt.2023.8 Published online by Cambridge University Press

Michie attached importance to Balfour's contribution as 'lay assessor'.[45] Michie did not invite anyone with an alternative academic background, despite non-technical writings on the topic having emerged prior to this time. Olson wrote that 'what the group badly needed in addition to computerologists was social anthropologists, political scientists, and philosophers'.[46]

The majority of Michie's panel were drawn from the highly interconnected 'AI establishment' and attached to top-ranking universities in the USA and the UK. Between the organizer and expert speakers, McCarthy was at MIT, Michie at Edinburgh and Bobrow participating in a Fulbright lectureship programme that had taken him from MIT to Edinburgh.[47] Further institutions represented on the panel were elite. Cordell Green joined from Stanford and John Robinson from Syracuse, and further invitees were again from MIT, Stanford and Edinburgh respectively in the cases of Papert, Nilsson and Meltzer, none of whom could attend.[48]

The Serbelloni conference nevertheless drew on a variety of participants from academia, industry and government. When compared with the participant list from McCarthy's Dartmouth Workshop of 1956, where AI was coined, for example, it is clear that Michie made substantial attempts to ensure that discussion at Villa Serbelloni was not purely academic.[49] Michie emphasized this variation in his participant biographies.[50] Bobrow was attributed with an 'unusual degree of eminence in the academic and the industrial worlds', and Peter Will was described as an expert in 'industrial electronics' and 'industrial robots'. Robert Taylor spent 'five years with the Advance Research Projects Agency in the Office of the Secretary of Defence' and Cordell Green had military experience with 'the U.S. Defence Department'. Professor Ito had both industry experience with Mitsubishi motors and government experience with the 'Japanese Ministry of International Trade and Industry'. Rather than incorporating experts on psychology, philosophy or sociology, this group were to draw on industry, academic, government and military experience. Michie's own view that this group was varied can perhaps be attributed to the narrow social web which constituted the AI 'establishment'.[51] Yet the final group already set the course of discussion such that the characterization of the problem by Gregory as ethical and therefore unquantifiable was unlikely to inform proceedings.

Conference proceedings

Debate at Villa Serbelloni reveals that, despite instances of self-criticism, participants largely refused to contemplate whether concerns over social implications should lead to limitations on their research. On the one hand, they actively identified risks which could arise due to their research. On the other hand, they concentrated on future forecasts rather than present action. At the close of this event, they collectively signed a reserved conclusion which conspicuously lacks any commitment to future work in this area.

[45] Rockefeller Foundation, op. cit. (2).

[46] Olson, op. cit. (44).

[47] Rockefeller Foundation, op. cit. (2).

[48] Donald Michie, correspondence with Miss Jane Allen on Serbelloni invites, 16 June 1971, Rockefeller Foundation Collection, Machine Intelligence Conference, RF RG 1.3, Series 120, Box 40, Folder 243, available via Rockefeller Archive Center, New York.

[49] The ten official participants for the Dartmouth workshop were Gerlertner, McCarthy, Minsky, More, Newell, Rochester, Samuel, Selfridge, Simon and Solomonoff. Ronald Kline, 'Cybernetics, automata studies and the Dartmouth Conference on Artificial Intelligence', *IEEE Annals of the History of Computing* (2010) 33(4), pp. 5–16.

[50] Rockefeller Foundation, op. cit. (2).

[51] Fleck, op. cit. (5)

https://doi.org/10.1017/bjt.2023.8 Published online by Cambridge University Press

Speculation and forecasts

The Serbelloni group's rhetoric reveals concerns among the group surrounding whether 1972 was the right time to consider the social implications of machine intelligence. McCarthy asserted that it was 'too early to speculate'.[52] Balfour admitted he was unsure 'the time is yet right to embark on any such exercise', while Olson described similar concerns among the group.[53] However, the discourse which took place on Lake Como reveals that a number of more complex assumptions and predictions led researchers to claim that these ethical problems could be delayed for a future date. And, while they may have thought it was too soon to act, their debate certainly suggests that they did not consider it 'too early to speculate'.

The Serbelloni group worked comfortably through the lens of future predictions. Michie presented participants with two copies of a predictive survey, one for the first day and one for the last day. He asked participants to place machines with 'intelligence approximating that of adult humans' and 'significant industrial spin-off' on a scale from five to fifty years. He asked whether machine intelligence would result in societal 'atrophy' or alternatively in 'enhanced' or 'unaffected' human intellectual and cultural life, and whether the 'risk of an ultimate "take-over"' was significant.[54]

Furthermore, despite the group expressing a wariness surrounding the topic of this debate, Cordell Green's unpublished conference outline demonstrates that the Serbelloni group did not consider it too soon to identify specific risks. They discussed the possible 'loss of freedom through transfer of decision-taking from people to machines' and 'loss of control through inability to understand complex systems'. Additional risks predicted include the possible 'disregard of human values by autonomous urban control networks', and the possibility of machine intelligence providing 'aids for political megalomaniacs' or 'facilities for democratic tyranny through enlarged techniques of social detection, persuasion and coercion'. Finally, they feared 'international cut-throat competition in machine intelligence'.[55] Not only did the Serbelloni group engage in self-criticism here, but they did so with significant accuracy.

Nevertheless, Michie made forecasts and quantified his predictions in a way which was typical of an 'AI establishment' whose overoptimistic projections have been identified as a cause of funding periodically drying up in what have since been simplistically named 'AI winters'.[56] Michie's publication of the results of the same survey he used at Villa Serbelloni, distributed on this later occasion to sixty-seven 'British and American computer scientists', reveals a desire to apply 'objective' methods of prediction to social implications.[57] Michie quantified results, presenting a graph to indicate predictions for the next fifty years. Michie acknowledged the limitations of this study but proposed in the future 'to find some objective basis of predicting the rate of development and social impact of machine intelligence'. Serbelloni participants did therefore question the promise of their research, but they did so through the lens of forecasting. This preoccupation with quantified prediction can be contrasted with the approach which 'social

[52] Boden, op. cit. (15), p. 164.

[53] Robert Balfour, correspondence with Donald Michie, June 1972, Donald Michie Collection, MS 88958/1/458, available via British Library, London; Olson, op. cit. (44).

[54] Donald Michie, machine intelligence survey, 1972, Rockefeller Foundation Collection, Machine Intelligence Conference, RF RG 1.3, Series 120, Box 40, Folder 243, available via Rockefeller Archive Centre, New York.

[55] Cordell C. Green, outline scheme of topics for social implications conference, 1972, Donald Michie Collection, MS 88958/1/458, available via British Library: London.

[56] Crevier, op. cit. (11), p. 203.

[57] Donald Michie, 'Machines and the theory of intelligence', *Nature* (London) (1973) 241(5391), p. 507.

https://doi.org/10.1017/bjt.2023.8 Published online by Cambridge University Press

anthropologists, political scientists, and philosophers' might have taken.[58] It also reveals parallels with approaches to AI risks which focus on extreme outcomes and often use surveys of the future as evidence that this focus on advanced AI capabilities is warranted.[59]

Contextualizing caution

To some extent, the Serbelloni group did devise mitigating strategies. In the case of computer use by megalomaniacs, Green described how 'countermeasures might include state monitoring and control of access to large database systems' and the 'identification of critical links in society at which monitoring could be exercised'.[60] In November of the same year, Michie once again recommended 'auditing procedures for computer programs' or 'programs to teach the users of intelligent systems'.[61] In these cases, the Serbelloni group acknowledged a possible role for government intervention in limiting the scope of machine intelligence projects.[62]

However, they largely focused on technological rather than political strategies, a practice which continues to persist in the field of AI ethics.[63] More 'research on program-understanding-programs' and on 'system-understanding-systems' was proposed to address the loss of control of AI. Their inclusion of predictions as extreme as people 'merging' with intelligent computing systems to form 'mixed societies', not alongside the risks, but instead as 'controversial as to whether good or bad', reveals the extent to which they prioritized unrestricted research.[64] This group's willingness to identify risks can therefore be contrasted with their unwillingness to act to change the course of their research.

In McCarthy's case, the claim that it was 'too early to speculate' was heavily informed by his own optimistic views about future technical capabilities, themselves arguably speculative given the lack of clear successes in machine intelligence capabilities at this time. In his planned speech for Villa Serbelloni, McCarthy wrote, 'with AI we will understand the consequences of alternate policies much better than we understand them now'.[65] This is a theme he repeated.[66] His willingness to speculate about risks but not to advocate for solutions was founded on his own confident speculations about machine intelligence research. McCarthy was not alone in taking this position. Green incorporated related propositions that machine intelligence itself may be able to address its own risks through 'computer-driven filtering' and 'validation of new policy proposals' when he wrote up the contents of the Serbelloni conference.[67]

Michie's cautious approach to present action was distinct from McCarthy's and, I argue, influenced more by practical circumstances than by ideological speculation. In particular, Michie was responding to events preceding the publication of *Artificial Intelligence: A General Survey*, the government-backed 1973 assessment on UK AI research which has

[58] Olson, op. cit. (44).

[59] Vincent C. Müller and Nick Bostrom, 'Future progress in artificial intelligence: a survey of expert opinion', in Vincent C. Müller (ed.), *Fundamental Issues of Artificial Intelligence*, Cham: Springer International Publishing, 2016, pp. 555–72.

[60] Green, op. cit. (55).

[61] Michie, op. cit. (57), p. 507.

[62] Michie, op. cit. (57), p. 507.

[63] Merve Hickok, 'Lessons learned from AI ethics principles for future actions', *AI and Ethics* (2021) 1, pp. 41–7.

[64] Green, op. cit. (55).

[65] John McCarthy, planned speech for Villa Serbelloni, 21 April 1972, Rockefeller Foundation Collection, Machine Intelligence Conference, RF RG 1.3, Series 120, Box 40, Folder 243, available via Rockefeller Archive Center, New York.

[66] McCarthy, op. cit. (65).

[67] Green, op. cit. (55).

https://doi.org/10.1017/bjt.2023.8 Published online by Cambridge University Press

since been called the Lighthill report. The applied mathematician and Lucasian Professor of Mathematics at the University of Cambridge, James Lighthill, called for AI funding to be directed away from ‘building robots’ and towards biomedical and industrial applications. Michie’s research was especially criticized, and this landmark paper had cascading impacts on his research funding.[68] Agar argues that this compelled Michie ‘to reflect on the extent to which his science did or did not, should or should not, respond to practical problems’.[69] While the final report was not published until the year following the Serbelloni conference, it was submitted in 1972. By the time Michie set the final agenda for the Serbelloni conference, he knew that this report would imminently shift funding away from his own research towards more applied projects. His correspondence reveals that he was defiant against this. He wrote in February that, ‘in short, James Lighthill should either (1) be for us, or (2) get with it’.[70] He made attempts to put Lighthill in touch with American machine intelligence researchers and to argue for the merit of his own research at Edinburgh.[71] During the lead-up to the Serbelloni conference, the supportive environment at Edinburgh came under significant strain, shifting Michie’s organizational efforts towards another priority: the defence of his own discipline.

As the events of the Lighthill debate unfolded, the scope of Michie’s conference can be seen to evolve through his communications with the Rockefeller Foundation. In April 1971, Michie planned to address ‘some very startling social repercussions’ and intended to title this retreat ‘the social responsibility of the machine intelligence researcher’.[72] He summarized the scope for debate: ‘the possible threats to man which might arise – displacement from employment, undermining the human intellectual self-image, military dangers etc’. Within these early communications, the Serbelloni Conference was framed entirely around risks posed by machine intelligence, with no suggestion that it would be necessary to identify practical benefits as well.

Nevertheless, reports written following the conference reveal that the scope had expanded. In addition to threats, the Serbelloni panel’s ‘terms of reference’ would now include benefits. They were to ‘determine some relevant categories of complex information systems and their applications’ and to discuss applications of machine intelligence both ‘good’ and ‘bad’.[73] Green’s 1972 report contained a rough balance of benefits and risks. Included in the Serbelloni group’s list were the ‘enlargement of the mental life of the ordinary citizen’; ‘powerful and perceptive computer aids for the artist, composer, writer’; and more ‘computer-based education and coaching in cultural awareness’ and ‘major (overwhelming) contribution to scientific and other understanding of our universe’.[74] This shift, occurring in the context of significant threats to Michie’s funding, indicates that the Lighthill report might have influenced Michie’s priorities surrounding this discussion. This shift towards the identification of industrial applications for machine intelligence may also be seen as part of wider trends within the scientific establishment in both the UK and the US where government funding for foundational research was drying up, with academics increasingly turning to the private sector.[75]

[68] James Lighthill, ‘Artificial intelligence: a general survey’, published as part of Science Research Council, Artificial Intelligence: A Paper Symposium, London: SRC, 1973, pp. 1–21.

[69] Jon Agar, ‘What is science for? The Lighthill report on artificial intelligence reinterpreted’, *BJHS* (2020) 53 (3), pp. 289–310.

[70] Donald Michie, correspondence on the Lighthill report, 21 February 1972, Donald Michie Collection, MS 88958/3/216, available via British Library, London.

[71] Agar, op. cit. (69).

[72] Michie, op. cit. (6).

[73] Green, op. cit. (55).

[74] Green, op. cit. (55), original emphasis.

[75] Philip Mirowski, *Science-Mart: Privatizing American Science*, Cambridge, MA: Harvard University Press, 2011.

A conclusion which was signed collectively by the Serbelloni group and shared with Rockefeller Foundation funders further reveals the contradictions encompassed by their approach. On the one hand, it emphasized that 'a wide range of increasingly important social consequences can now be seen to be following developments in the field of computer science and automation'. On the other hand, it did not advocate for action. Instead, it articulated that more time was needed for machine intelligence to be 'developed to any significant degree'.[76] This limited conclusion was informed both by faith in their own research providing policy recommendations, and by concern that their own funding might be under threat. For individual participants, further research will be needed to examine exactly how these two factors interacted to influence their world view.

The aftermath

Of all the Serbelloni participants, Michie demonstrated the most concerted interest in the social implications of his research. I have uncovered four key attempts to extend interest in this topic. First, he proposed a two-year study on the social implications of machine intelligence in consultation with the Serbelloni group.[77] Second, he made numerous attempts to organize a second conference on societal impact in 1974, 1975 and 1976.[78] Third, in 1977 he brought up social implications at a Machine Intelligence Workshop, this time in Leningrad.[79] Finally, prior to 1984, he proposed that Edinburgh University teach undergraduates about social implications.[80] Despite Michie's position as an organizational leader within the social web of machine intelligence, he was unable to secure continued interest from within his community and so, in the years following 1972, his approach began to bring in other elite and powerful thinkers, in particular from the humanities.

Donald Michie: a discontent's approach to technological determinism

Following the Serbelloni conference, Michie attempted to reconvene the group for a longer study. He did so despite the muted conclusion signed by the working party and against the advice of Balfour, who wrote that there 'seemed rather little support' for this idea from other members of the working party.[81] Michie nevertheless requested four thousand pounds for a 'two-year study of social implications of research in machine intelligence'. He planned to consult the same working party, demonstrating a prolonged effort to involve the 'AI establishment' in these discussions.[82] Additionally, he aimed to consult 'additional expert witnesses' to address 'educational and socio-political considerations'.[83] He admitted that these topics were only treated 'superficially' at Villa Serbelloni.

Michie attempted to convene the same group three more times in Spetsai, Greece. Due to difficulties securing funding and participants, he failed in scheduling this conference

[76] Michie, op. cit. (6).

[77] Michie, op. cit. (23).

[78] Donald Michie, correspondence with Balfour on Spetsai conference, April 1975, Donald Michie Collection, MS 88958/1/458, available via British Library, London.

[79] Donald Michie, correspondence on the Leningrad conference, May 1977, Donald Michie Collection, MS 88958/1/458, available via British Library, London; Robert Balfour, correspondence on the Leningrad conference, May 1977, Donald Michie Collection, MS 88958/1/458, available via British Library, London.

[80] Donald Michie, 'Machine intelligence: philosophy, social implications and practice' (197?), Donald Michie Collection, MS 88958/1/298, available via British Library, London.

[81] Michie, op. cit. (78).

[82] Michie, op. cit. (23).

[83] Michie, op. cit. (23).

https://doi.org/10.1017/bjt.2023.8 Published online by Cambridge University Press

for 1974, 1975 or 1976.[84] Michie's approach during these years marked a split between his interest in the social implications and his involvement with the machine intelligence community more generally. By this stage, his primary correspondent on social implications was Balfour, the only Serbelloni participant with a non-AI background, and these letters show Michie's frustration at the lack of prolonged interest in social implications from the remaining Serbelloni participants. Michie wrote of the necessity of 'broadening influences' and invited experts from the humanities to Spetsai. Not only was Michael Hurst, a fellow in history and politics from St John's, Oxford, invited to Spetsai, but also Michie proposed that if a book was written on social implications, Hurst would be his candidate author.[85] Michie prioritized humanities scholars over the machine intelligence community when he accommodated his 'broadening influences' by severing the planned Spetsai conference from the larger Machine Intelligence Workshop. This was indicative of Michie's desire to take a broader approach to social implications in the mid-1970s but also in part resulted from the lack of interest expressed in the machine intelligence community, as is indicated in Balfour's correspondence with Michie.

Another of the 'broadening influences' consulted by Michie during this period reveals a political dimension to his work on social implications. Michie left the Communist Party in the 1950s. Nevertheless, his commitment, alongside that of his first wife, biologist Anne McClaren, to socialism and activism has been described as lifelong.[86] His son-in-law writes, 'from the world peace congresses of the 1950s to the recent anti-war demonstrations they were always there, often together'.[87] His daughter, Susan Michie, similarly describes a lifelong interest in political theory and Marxism in particular.[88] Michie frequently incorporated this socialist world view in his scientific writing, spending years as science correspondent for the *Daily Worker*.[89] It is therefore unsurprising that when Michie attempted to involve political figures in his work on social implications, he turned to the Labour Party. He invited Shirley Williams, from 1974 Secretary of State for Prices and Consumer Protection and from 1976 to 1979 Secretary of State for Education, to the Spetsai conference and continued to involve her in his work on machine intelligence.[90] Eventually he appointed her to the board of the Turing Institute, an AI laboratory set up by Michie in Glasgow in 1983.[91] It is also notable that one other attempt to discuss 'social implications' with the machine intelligence community saw Michie engage not with prominent US researchers, but rather with the USSR, as he advocated for 'East–West collaboration' at the Leningrad Machine Intelligence conference of 1977 and wrote to Balfour of the importance of collaborating on 'neutral soil'.[92]

In addition to these initiatives, between 1973 and 1984, Michie proposed an Edinburgh University course on 'machine intelligence: philosophy, social implications and practice'.[93] He included a segment dedicated to 'social aspects of AI'. In this course, Michie planned to focus on predicting 'rates of development' in machine intelligence.[94] As preliminary reading he recommended only one text, 'Forecasting and assessing the impact of artificial intelligence on society', itself an output from another 'establishment' AI conference – the

[84] Michie, op. cit. (78).
[85] Michie, op. cit. (78).
[86] Michie, op. cit. (23).
[87] Michie, op. cit. (17), p. 257.
[88] Susan Michie, discussion on the life of Donald Michie, phone interview, 13 May 2021.
[89] Michie, op. cit. (17), p. 257.
[90] Michie, op. cit. (78).
[91] Shirley Williams, *Climbing the Bookshelves*, London: Virago, 2009.
[92] Michie, op. cit. (79).
[93] Michie, op. cit. (80).
[94] Michie, op. cit. (80).

https://doi.org/10.1017/bjt.2023.8 Published online by Cambridge University Press

International Joint Conference on AI.[95] This took a systematic approach to surveying researchers and consequently predicting future applications of machine intelligence. It closely resembled Michie's own 1972 survey. The paper also articulated a 'strong mood of optimism' and Michie's recommendation of it reveals an approach still focused on forecasting and on the quantification of societal questions. In line with this approach of politically informed preparation for an inevitably technological future, Michie opposed arguments presented in Weizenbaum's *Computer Power and Human Reason*, primarily because he considered that science was ultimately the way forward.[96] Michie wrote that 'the committed scientist who believes, as I do, that reason, general increase of knowledge and technological extension of human powers are, on balance, forces for good will react against Weizenbaum's demagogy'.

After the aftermath: apparent polarities amidst establishment ties

Looking to the Serbelloni conference and subsequent initiatives by Donald Michie, numerous differences within the AI 'establishment's' approach to social implications become clear. They disagreed on how to respond to the risks posed by machine intelligence. Gregory proposed that machine 'intelligence' should be substantially different from human 'intelligence' to avoid the worst impacts on society, while Bobrow proposed government control over large data sets to limit their use by 'political megalomaniacs'.[97] McCarthy advocated for waiting, as AI was the best hope of 'objective' solutions to societal problems, and Michie advocated for increased education and debate surrounding social implications.[98] He recommended mitigating strategies, including algorithmic auditing and obligations to politicians.

However, there was much which tied this group's approach to social implications together as an AI 'establishment'. Members of the Serbelloni group agreed that identification of risks was important. There was also complicated rather than binary disagreement over what should be done in the face of these risks, but an approach focused on forecasting and on international research collaboration was held in common by the group.

Indeed, the Serbelloni group and Weizenbaum also had much in common. First, they all belonged to the overlapping social group based at elite academic establishments. Second, their ideas on the social implications of machine intelligence overlapped. Both focused on problems surrounding incomprehensible programmes, acknowledged the importance of looking to side effects of machine intelligence as the primary site for significant societal harm, and articulated significant concern surrounding the impact their research may have on man's image of himself, described by this group using highly gendered language. Despite significant differences, in particular in their faith in the machine intelligence research programme, Weizenbaum and Michie were not polar opposites, but rather both important contributors to 1970s debate on the social implications of machine intelligence, both influenced by socialist politics.

As time went on, many of the predictions made by this group of elite researchers came to pass. Society has since grappled with the moral dilemmas posed by legal AI, where

[95] Oscar Firschein, Martin A. Fischler, L. Stephen Coles and Jay M. Tenenbaum, 'Forecasting and assessing the impact of artificial intelligence on society', in *Proceedings of the 3rd International Joint Conference on Artificial Intelligence* (Stanford, 20–3 August 1973), Stanford: International Joint Conferences on Artificial Intelligence Organization, pp. 105–20.

[96] Michie, op. cit. (10), pp. 743–5.

[97] Green, op. cit. (55); Gregory, op. cit. (3), pp. 3–13.

[98] John McCarthy, 'Technology and the enhancement of man', June 1973, Stanford Digital Repository, John McCarthy Papers, book proposal, Box 3, Folder 4, Stanford, Green Library, at https://purl.stanford.edu/mt301nd7838 (accessed 5 May 2021); Michie, op. cit. (78).

https://doi.org/10.1017/bjt.2023.8 Published online by Cambridge University Press

algorithms are used to predict recidivism rates, transforming and displacing the human judgement demonstrated by judges.[99] Cut-throat competition in AI has caused significant controversy.[100] And incomprehensible computer programs have led recent initiatives to prioritize algorithmic 'explainability'.[101] However, successful predictions cannot be attributed to this 'AI establishment' based on technical expertise. Following the 1970s, the symbolic AI approach was largely abandoned in favour of neural networks, and gradually the very harms predicted by the Serbelloni group came to pass because of new methods which they had not considered.[102] Yet perhaps the accuracy of these predictions can shed light on the continuation of so many of the power structures that emerged as part of the military–industrial–university complex at this time. As discussed by Giroux in *University in Chains*, the government and corporate influences on scientific research which emerged in the twentieth century continue to this day, making it unsurprising that whatever form of machine intelligence emerged, there would be 'international cut-throat competition' alongside risks of it being used for 'social detection, persuasion and coercion'.[103]

Conclusion

Through the 1970s, leading AI researchers played an important role in initiating debate about the social implications of intelligent machines. The archival materials I have explored, detailing the Serbelloni conference and subsequent initiatives of Donald Michie, reveal the extent to which this group took an approach to societal impact that was characteristic of their 'establishment'. They prioritized funding, used quantitative predictive methods, and ultimately embraced the future of science and machine intelligence research. However, the Serbelloni group simultaneously questioned the future that they anticipated creating through symbolic AI. They identified specific societal risks of machine intelligence, discussed their obligations to society, and proposed mitigating strategies.

In this paper, I have uncovered significant variation within the Serbelloni group. Gregory, Michie and McCarthy each envisioned looming social implications in distinct ways that cannot easily be summarized under the labels of *pioneer* or *discontent*. They each utilized the status that elite university professorships brought them to shape society. The material I have unearthed shows that they all did so in a way which acknowledged ethical problems raised by their research. As a result, they shaped 1970s debates on solutions to these problems and influenced subsequent generations. Yet grouping Michie together with the remainder of the Serbelloni group would be misleading and his contribution to the discourse on the social implications of machine intelligence should be acknowledged alongside his more widely cited role as a staunch defender of AI.

Acknowledgments. I am grateful to a number of people who have been integral to the writing of this paper since 2020. I would like to thank the editors of this special issue and organizers of the Mellon Sawyer Seminar on

[99] Ziyaad Bhorat, 'Do we still need human judges in the age of artificial intelligence?', Open Democracy (8 August 2017), at www.opendemocracy.net/en/transformation/do-we-still-need-human-judges-in-age-of-artificial-intelligence (accessed 1 June 2021).

[100] Mariarosaria Taddeo and Luciano Floridi, 'Regulate artificial intelligence to avert cyber arms race comment', *Nature* (London) (2018) 556(7701), pp. 296–8.

[101] Thilo Hagendorff, 'The ethics of AI ethics: an evaluation of guidelines', *Minds and Machine* (Dordrecht) (2020) 30(1), pp. 99–120.

[102] Boden, op. cit. (15), p. 9.

[103] Henry Giroux, *University in Chains: Confronting the Military–Industrial–Academic Complex*, New York: Routledge, 2007.

https://doi.org/10.1017/bjt.2023.8 Published online by Cambridge University Press

Histories of AI: Richard Staley, Sarah Dillon, Syed Mustafa Ali, Stephanie Dick, Matthew Jones and Jonnie Penn. In particular I would like to thank Jonnie Penn for his role in guiding me through this topic from the outset, providing direction and invaluable advice throughout. Finally, I am grateful for all opportunities to workshop this paper with participants in this special issue of the *BJHS – Themes*: *Histories of AI: A Genealogy of Power*.

Cite this article: Powell R (2023). The 'artificial intelligentsia' and its discontents: an exploration of 1970s attitudes to the 'social responsibility of the machine intelligence worker'. *BJHS Themes* **8**, 189–203. https://doi.org/10.1017/bjt.2023.8