Mind and Machines

Session 1: The Philosophy of Palantir

Optional Pre-Reading

Index

News

1.	**Andy Burnham searches for UK alternative to Palantir** Financial Times · 1 October 2026	4
2.	**Patient privacy fears as US spy tech firm Palantir wins £330m NHS contract** The Guardian · 21 November 2023	8
3.	**In defence of Palantir** The Telegraph (Matthew Lesh) · 4 May 2026	10
4.	**Palantir's access to identifiable NHS England patient data is 'dangerous', MPs say** The Guardian · 11 May 2026	12
5.	**MPs warn that Palantir's increasing presence in the UK public sector is an "unacceptable point of weakness"** UK Parliament, Science, Innovation and Technology Committee · 3 June 2026	14
6.	**MPs urge Labour to ditch £330m Palantir software contract with NHS** The Guardian · 9 July 2026	16
7.	**Palantir boss: MPs who want to block us from NHS are 'noisy minority'** The Telegraph · 29 July 2026	18
8.	**Correcting the Record: Response to the EFF January 15, 2026 Report on Palantir** Palantir Blog · 27 January 2026	20
9.	**Palantir Has a Human Rights Policy. Its ICE Work Tells a Different Story** Electronic Frontier Foundation · 20 April 2026	23
10.	**ICE Is Paying Palantir $30 Million to Build 'ImmigrationOS' Surveillance Platform** WIRED · 18 April 2025	26
11.	**Palantir lands $10 billion Army software and data contract** CNBC · 1 August 2025	28
12.	**Palantir's Billionaire CEO Just Can't Stop Talking About Killing People** Gizmodo · 7 February 2025	29
13.	**Thiel's Palantir, Israel Agree Strategic Partnership for Battle Tech** Bloomberg (reprint) · 12 January 2024	30
14.	**Israel is using artificial intelligence to help pick bombing targets in Gaza, report says** CNN (via KION) · 3 April 2024	31
15.	**We work for Google. Our employer shouldn't be in the business of war** The Guardian (open letter signed by Google employees) · 4 April 2018	33
16.	**Former Palantir workers condemn company's work with Trump administration** NPR · 5 May 2025	34
17.	**Ukraine is using Palantir's software for 'targeting,' CEO says** Reuters · 2 February 2023	36
18.	**'Lavender': The AI machine directing Israel's bombing spree in Gaza** +972 Magazine and Local Call · 3 April 2024	36

Philosophy

19.	**Killer Robots** Robert Sparrow, Journal of Applied Philosophy 24(1): 62–77 · 2007	38
20.	**The responsibility gap: Ascribing responsibility for the actions of learning automata** Andreas Matthias, Ethics and Information Technology 6: 175–183 · 2004	52

Law

21.	**Principles of International Law Recognized in the Charter of the Nürnberg Tribunal and in the Judgment of the Tribunal** United Nations, International Law Commission · 1950	62
22.	**Protocol Additional to the Geneva Conventions (Protocol I): Articles 36 and 57** International Committee of the Red Cross, IHL Databases · 8 June 1977	64
23.	**UK GDPR: Article 4(7)–(8) and Article 28 (Processor)** legislation.gov.uk · 2016/679, as retained in UK law	65

News

Andy Burnham searches for UK alternative to Palantir

Financial Times · 1 October 2026

https://www.ft.com/content/9a6e17fe-55e4-4a6d-8761-8346ec765ba9

Palantir Technologies Inc

Andy Burnham searches for UK alternative to Palantir

Prime minister's 'Buy British' push triggers hunt to find replacement for controversial US tech group awarded health service and defence contracts



Palantir has come under fire over its links to the Trump administration and the US defence sector © Hannibal Hansche/EPA/Shutterstock

Chris Smyth, **Lucy Fisher** and **Laura Hughes** in London

Published OCT 1 2026 | Updated OCT 1 2026, 23:26

Unlock the Editor's Digest for free

Roula Khalaf, Editor of the FT, selects her favourite stories in this weekly newsletter.

Enter your email address

Sign up

Andy Burnham's government is searching for a UK alternative to Palantir, as the prime minister's "Buy British" push triggers a hunt to find a homegrown replacement for the controversial US software group in the NHS and the Ministry of Defence.

Ministers have identified a handful of UK technology companies they hope might be candidates to take over the health service's Federated Data Platform, as they consider applying a break clause in Palantir's contract in March next year, people familiar with the matter told the FT.

Meanwhile, a defence minister has told parliament that the government is sounding out alternative suppliers for its MoD contract.

Some government figures are pushing to replace Palantir, which has come under fire over its links to the Trump administration and the US defence sector, to reduce the risk of a backlash by progressives against Burnham's drive to foster the British AI industry.

Palantir won a £330mn contract from NHS England (NHSE) in 2023 to run the health service's Federated Data Platform, a system designed to drive efficiencies and improve care in hospitals by joining up data better.

But the project has been beset by controversy, with NHSE claims about the success of the platform under growing scrutiny after the FT identified a litany of errors in a key underlying dataset.

This year, a minister was forced to apologise after the FT revealed some Palantir staff would have "unlimited access" to identifiable patient information while working on one part of the FDP, which NHSE had failed to disclose.

One option under consideration is to offer Palantir a temporary contract extension of about 12 months to give more time to develop a British alternative, according to people familiar with discussions. The current contract stipulates a minimum two-year extension from March.

Ministers are also considering whether the FDP programme could be divided between several companies, rather than run solely by Palantir.

Quantexa, which was in the running for the FDP contract in 2023, is among three or four British tech companies being considered, according to people familiar with discussions. The others are not known.

However, no decisions have yet been taken, and some Labour figures argue that Palantir's UK-based business should be considered a British company under Burnham's definition.



Louise Haigh, first secretary of state, said the government wanted to build 'sovereign capabilities' in AI © Ian Forsyth/Getty Images

Tom Bartlett, former deputy director of data engineering and integration at NHS England, said it would be "hugely risky" to replace Palantir and "a leap into the unknown".

"There isn't a platform that can replace the functionality that Palantir can bring," he said. "They [the NHS and government] can assemble a group of people who are confident they could do it, but they would be proving themselves for the first time and it would be a gamble."

While the government began a review of Palantir's contract when Sir Keir Starmer was prime minister, a specific search for a British alternative is under way after Burnham said he wanted more government spending to go to UK businesses.

“I don’t think we should allow the culture of sending everything outside of this country to continue — this will be a big change under this government,” he said in his first House of Commons appearance as prime minister.

At the Labour Party conference in Liverpool this week, first secretary of state Louise Haigh said the government wanted to build “sovereign capabilities” in AI and boost the UK tech sector. This would ensure the country would not “be in hock to foreign giants”, she added.

One Labour figure confirmed that Haigh’s intervention was “a shot across the bows” at Palantir, adding: “There’s a wider question here about building up sovereign capabilities and capacity — and we do have world-leading UK expertise in this area.” The person said there was a general sense among officials that the UK state was “overly reliant on Palantir and overseas capabilities more widely”.

The Cabinet Office, which runs the machinery of government, has recently reviewed its public procurement rules and rewritten the “social value” requirements to focus more narrowly on UK job creation and skills development.



AI minister Kanishka Narayan: ‘The British public wants to make sure that their private data… isn’t in the hands of… foreign firms abroad’ © Mary Turner/Bloomberg

However, officials have stressed that these criteria are legally “country agnostic”: while they favour companies supporting British jobs and training, that does not have to mean British-owned groups.

Another government figure said there was a particular focus on sovereign capabilities in health technologies, which ministers hoped could transform the NHS while fostering an industry based on its rich data sets.

AI minister Kanishka Narayan told a Politico event on the sidelines of Labour’s conference that polling consistently showed “the British public wants to make sure that their private data, particularly in healthcare, isn’t in the hands of foreign billionaires and foreign firms abroad”.

James Frith, the health innovation minister, told MPs last month that he was “concerned about mistrust of Palantir and the impact it could have on people’s willingness to share data with the NHS”.

Questions about Palantir’s involvement in the health service are seen as more acute than its work with the MoD, according to officials, but it is also being examined.

In November 2022, Palantir was awarded a three-year contract by the MoD worth £75.2mn after the tender was not advertised. In December 2025, the company was directly awarded another three-year contract, again without any competition, worth £240.6mn.

Speaking in June, in response to questions on the contract, defence minister Lord Vernon Coaker said: “We are seeking to look at options with small and medium-sized and other companies that could provide the same capability, so that we have a competitive process in future.”

The process would “keep Palantir on its toes in the marketplace as well”, he added.

Other European governments are seeking to cut their reliance on American technology providers amid growing concern over US President Donald Trump’s foreign policy.

France and Germany have excluded or cut contracts with Palantir from their armed forces and intelligence agencies in recent months, seeking more local alternatives.

Palantir said in a statement that its software was “helping to deliver better care” and that it was “deeply committed to the UK”.

“More than one in six of our global workforce is based here . . . And all UK government data is stored in the UK,” it added.

The Department of Health did not immediately respond to a request for comment.

An MoD spokesperson said: “As set out in the Defence Investment Plan, we will bring forward a package of ‘backing British’ measures to ensure that defence spending is driving value back into local communities, while remaining consistent with our international obligations.

“We continue to work with a range of international suppliers based on operational requirements, including Palantir, which is a longstanding technology investor in the UK economy.”

Copyright The Financial Times Limited 2026. All rights reserved.

Patient privacy fears as US spy tech firm Palantir wins £330m NHS contract

The Guardian · 21 November 2023

https://www.theguardian.com/society/2023/nov/21/patient-privacy-fears-us-spy-tech-firm-palantir-wins-nhs-contract



The NHS has sparked controversy by handing the US spy tech company Palantir a £330m contract to create a huge new data platform, leading to privacy concerns around patients' medical details.

The move immediately prompted concerns about the security and privacy of patient medical records and the suitability of Palantir to be given access to and oversight of such sensitive material.

NHS England has given Palantir and four partners including Accenture a five-year contract to set up and operate the "federated data platform" (FDP).

Palantir is known for working closely with intelligence agencies and military organisations around the world, such as the CIA and UK Ministry of Defence. Peter Thiel, its billionaire founder and chair, backed Donald Trump in his 2016 presidential election.

Thiel, a libertarian, told an Oxford Union debate in January that the NHS makes people sick and should be privatised and Britons' attachment to it showed they were exhibiting "Stockholm syndrome".

The platform will be based on software that will allow individual health service trusts as well as the NHS's 42 integrated care systems – regional groupings of trusts – to "talk" to each other digitally and share data in an attempt to improve care, cut waiting times and make the service more efficient.

For example, it will bring together real-time data on the number of beds in a hospital, the size of waiting lists for planned care and staff rosters, to help hospitals do more operations.

But its creation has been overshadowed by Conservative, Labour and Liberal Democrat MPs, as well as tech, medical and civil liberties groups, voicing unease about Palantir, the potential for patients' data being mishandled and also whether patients will be able to opt out of the FDP sharing their data.

Amnesty International said Palantir's involvement in "serious human rights abuses" meant NHS England should have rejected the bid. The company has come under fire for helping the US immigration enforcement agency (ICE) implement a tough deportation policy that Trump introduced in 2017. Its technology helped enable ICE to identify, exchange information about and track migrants and asylum seekers, and also to plan raids that led to parents being arrested, separated from their children and deported, Open Democracy reported.

Palantir has also supplied technology to governments that allows them to spy on their citizens.

"Palantir is a very troubling choice of service provider for the NHS, given the human rights controversies surrounding the company", said Peter Frankental, Amnesty's business and human rights director. "Any NHS public procurement

tenderers whose activities have been linked to serious human rights abuses, as is the case with Palantir, should be excluded on grounds of 'grave professional misconduct', as permitted under procurement law."

The public needed "assurance that their personal information won't be harvested by Palantir for purposes that have little to do with their health", Frankental added.

The British Medical Association, which had previously voiced concern about the NHS's alleged lack of scrutiny of bidders on "ethical" grounds, said Palantir's winning bid was "deeply worrying".

NHS England sought to allay such concerns. It stressed that none of the companies in the winning consortium would be able to access health and care data without its explicit consent; that it would retain control of all data within the platform; and that it would not include GP data.

It said the new software would be protected by the highest possible standards of security through the deployment of "privacy enhancing technology".

It had said the day before the announcement that patients would not be allowed to opt out because it would anonymise all data before it was shared and also because the data would be used for "direct patient care", for which no opt-out is available. The Foxglove campaign group said it was ready to launch a legal challenge to the apparent denial of a right to opt out.

Sam Smith, of MedConfidential, said: "Whether or not patients will be able to opt out is among the most important of the many unanswered questions about this £330m data project that NHS England's announcement does not clarify."

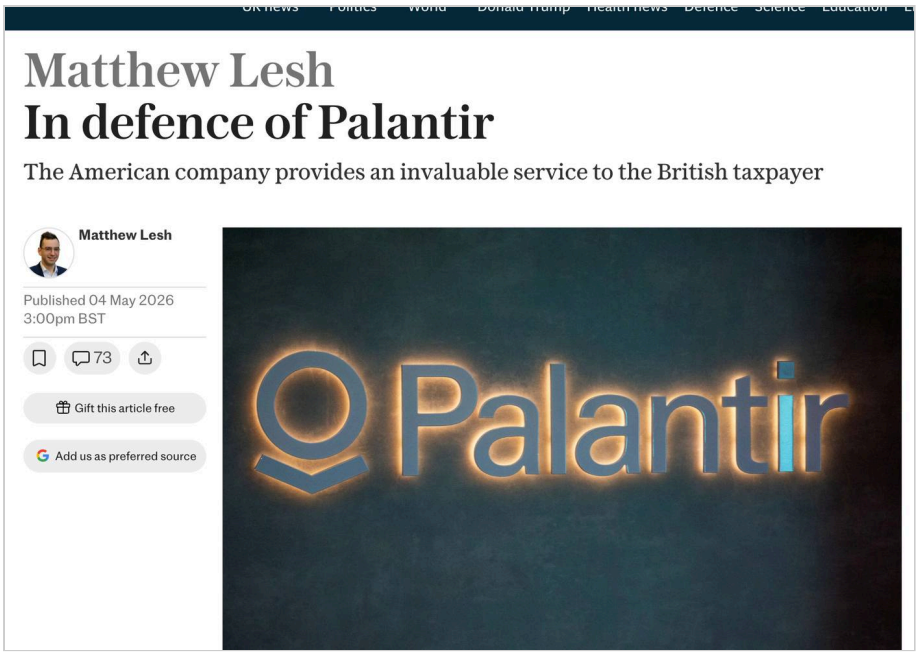
Matthew Taylor, the chief executive of the NHS Confederation, said the platform would be "game-changing". He said: "Clinicians will be able to access live data of available theatre slots, staff availability and individual patient data suitable for particular procedures at the touch of a button." It could also help the NHS improve how long patients wait for treatment, he added.

But he also warned that the NHS's unhappy previous history of trying to bring its huge stores of data together – opposition has forced it to abandon two similar projects since 2012 – might mean it struggles to convince the public of the FDP's potential benefits.

In defence of Palantir

The Telegraph (Matthew Lesh) · 4 May 2026

https://www.telegraph.co.uk/news/2026/05/04/in-defence-of-palantir-peter-thiel-left-nhs-digital-health/



The Left has a new bogeyman: Palantir, the American company that develops data integration and analytics software.

Palantir is responsible for the NHS's Federated Data Platform (FDP), which is connecting data across hospitals, trusts and care services and provides tools to improve planning, reduce waiting lists and help staff make better-informed decisions. It's hardly particularly exciting stuff, but it is important.

According to the NHS, the FDP has already improved scheduling efficiency in operating theatres, leading to 110,078 additional patients undergoing procedures. It has integrated hospital records with social care teams, reducing delayed discharges by 15.3 per cent. By linking cancer diagnostic, patient, and treatment systems, it has increased the rate of diagnosis within 28 days by 6.8 per cent. It has also reduced waiting lists by 797,728 patients.

This clear success has not stopped Zack Polanski from calling for Palantir to "pack its bags and get the hell out of the NHS". Nor did it prevent Faiza Shaheen of Tax Justice UK, sounding like an uninformed, petulant child on BBC Politics Live, arguing that the company cannot be trusted with patient data.

Polanski and Shaheen appear to have little to no idea how the system actually operates. To start, Palantir does not own, have access to, or control patient data. That remains entirely in the trusts' control. Just like using Microsoft Word does not mean that Microsoft owns your data. If the NHS were to stop using Palantir's services, it could migrate the data and any applications built on the platform to another service. If anything, patient data is significantly more secure than before, as FDP carefully controls access and replaces whiteboards in full view on wards and spreadsheets saved in obscure locations.

The critics also target Palantir because of its politics and clients. The company is unashamedly pro-Western and, unlike some in Silicon Valley, is committed to supporting hard power in a more dangerous world. It was co-founded by Peter Thiel, who has well-known conservative views and backed Donald Trump. It provides services to the American military and immigration enforcement (which begun under the Barack Obama administration), as well as the Israel Defence Forces. It also provides services to Ukraine's military, often overlooked by the company's critics, though perhaps Shaheen would prefer that Volodymyr Zelensky forgo the best available technology to defend his nation too.

For those actually involved with the NHS, none of this is even remotely relevant. NHS England's recently retired chief executive, Amanda Pritchard, told a parliamentary committee the system is "already paying back". Tom Bartlett, who until recently led the engineering team building the platform within the NHS and has decades of experience in similar projects, has said: "In 20 years of working in the NHS, nobody came up with anything like the FDP". Despite claims that Palantir could be easily ripped out and replaced with a British equivalent, Bartlett has made it clear that there is "no credible alternative" and that if Palantir were removed, the "NHS would go backwards, not forwards".

Last time the NHS tried to build its own technology at this scale, it was a disaster. The NHS National Programme for IT, a project with similar goals in the 2000s, cost taxpayers around £10bn-£12bn but was so dysfunctional that it was

ultimately abandoned. By contrast, Palantir's contract is worth £330m over 7 years – coming to just £200,000 per year per organisation. This is a minuscule cost compared to any imaginable alternative. The Treasury's infrastructure authority, NISTA, rates it green, a status achieved by only around 15 per cent of major government programmes. Over 123 trusts are already using it, and 80 have reported benefits.

FDP's current functionality, six products, is only just the beginning. Over time, it will become seamless for patients to move between the NHS and other services, with all caregivers across the system fully informed about patient needs. The NHS will soon begin to benefit from a marketplace of shared products built by a multitude of third-party developers. AI is also integrated across the system, so staff with little technical expertise will be able to develop their own applications and analytical outputs. This is like the iPhone, but in 2008. There is still so much more to come.

Ripping up all this would be an obvious, monumental mistake. It's hard to avoid the reality that the opponents of Palantir would prefer that patients die rather than use the best available piece of software. They have allowed a dumb and dangerous ideology to infect their brains. But rest assured, over time, this is more than likely to wash away. The patient benefits are just too colossal to ignore.

Matthew Lesh is Country Manager at Freshwater Strategy and a Public Policy Fellow at the Institute of Economic Affairs

Palantir’s access to identifiable NHS England patient data is ‘dangerous’, MPs say

The Guardian · 11 May 2026

https://www.theguardian.com/society/2026/may/11/palantir-access-nhs-england-patient-data



MPs have warned that an NHS decision to grant Palantir access to identifiable patient information in its plan to use AI to improve the health service is “dangerous” and will fuel public fears that data privacy is not being prioritised.

NHS England has allowed staff from the US tech firm and other contractors to access patient data before it has been pseudonymised, despite internal fears of a “risk of loss of public confidence”, the Financial Times reported.

The health service made the move to allow Palantir to access the data in recent weeks according to the reports, which revealed an internal NHS briefing that said it would allow “unlimited access to non-NHSE staff” to part of the NHS’s federated data platform (FDP), which holds identifiable patient information.

Palantir, which also supports Donald Trump’s ICE immigration crackdown and the Israeli, US and UK militaries, was awarded a £330m contract to help build the FDP, installing AI systems to integrate scattered health datasets and bring efficiencies to medical treatment. But the deal has been dogged by warnings from campaigners and MPs concerned about the security of patient records.

The Patients Association said it was concerned patients were not consulted on a significant change to who has unlimited access to patient data. Rachel Power, its chief executive, said patients wanted “transparency, clear boundaries around access to their data, and to be consulted when changes to those agreements are proposed”.

The leaked NHS England briefing acknowledged the “considerable public interest and concern about how much access to patient data Palantir/Palantir staff have”. In 2023, shortly after the deal was agreed, NHS England said it would ensure “personal data remains protected and within the NHS at all times”.

NHS England stressed external consultants requiring data access must have government security clearance and that it had “strict policies in place for managing access to patient data”. With hundreds of different datasets in the FDP system, it was becoming time-consuming for contractors, including Palantir engineers, to apply for individual permissions. Instances when they did see identifiable patient data while working on the system’s “pipelines” were logged. They did not have permission to remove the data from the NHS, Palantir said.

But the MP Rachael Maskell, a former NHS worker who is calling for the Palantir project to be stopped, said: “As Palantir get their claws deeper into our NHS data we can see how it is opening it up to greater private interest. This is a dangerous development and I ask the government to get a grip on this project before it is too late.”

Palantir said it was a “data processor” and not a “data controller”, meaning its software could only be used to process data precisely in line with customer instructions. “Using the data for anything else would not only be illegal but technically impossible due to granular access controls overseen by the NHS,” it said.

Martin Wrigley, a Liberal Democrat member of the Commons technology select committee, said of the NHS move: “This somewhat cavalier attitude to data security demonstrated how this whole project does not have security by design at its heart. The public will be rightfully concerned that data privacy is not the first concern.”

Palantir is facing opposition to its widening role in the UK public sector. Last month the Guardian revealed the company was closing in on a deal to widen its work with the Metropolitan police to use AI to analyse intelligence in criminal investigations, while hundreds of thousands of citizens and numerous backbench MPs oppose its role.

Polling last week showed more than two-thirds of the UK public are concerned at Palantir’s growing number of public contracts and 40% distrust it to not access NHS patient data, despite the company repeatedly saying it cannot and will not do so.

Tom Hegarty, the head of communications at Foxglove, a tech equity campaign group, said: “NHS patients never consented to have their data accessed by a company like Palantir whose record is in targeting people, not caring for them … Once again: Palantir fails the trust test. The government should … cut Palantir out of our NHS once and for all.”

NHS England said access would be given to a small number of people working on the new data collection platform.

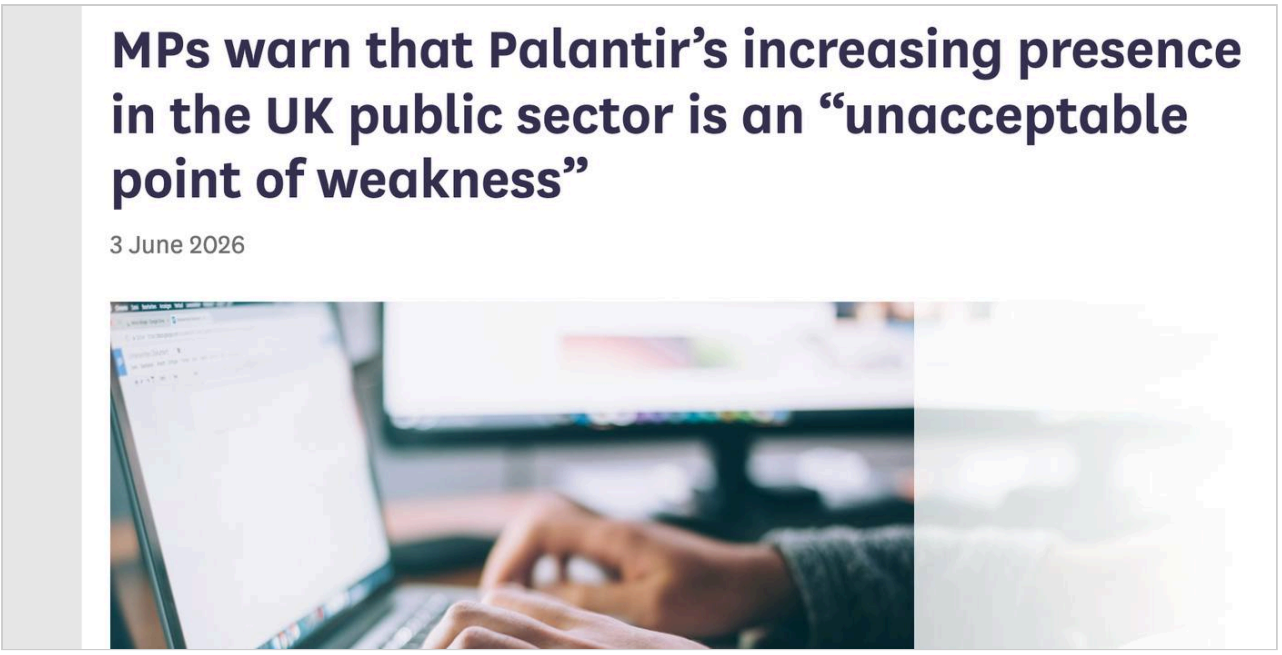
A spokesperson said: “The NHS has strict policies in place for managing access to patient data and carries out regular audits to ensure compliance – including monitoring the work of engineers helping to set up the central data collection platform that will track NHS performance and help improve care for patients.

“Anyone external requiring access must have government security clearance and be approved by a member of NHS England staff at director level or above.”

MPs warn that Palantir’s increasing presence in the UK public sector is an “unacceptable point of weakness”

UK Parliament, Science, Innovation and Technology Committee · 3 June 2026

https://committees.parliament.uk/committee/135/science-innovation-and-technology-committee/news/214048/mps-warn-that-palantirs-increasing-presence-in-the-uk-public-sector-is-an-unacceptable-point-of-weakness/



Palantir shouldn’t play such a significant role in the UK public sector, a new report by the Science, Innovation and Technology Committee warns, stressing that it’s not the only company capable of providing the ‘middleware’ required by public bodies.

The cross-party committee identifies Palantir as the most concerning example of the public sector’s growing reliance on a small number of major technology providers, including Microsoft and Amazon Web Services.

MPs urge the government to exercise the 2027 break clause in the NHS Federated Data Platform Contract with Palantir and either develop an in-house replacement or seek an alternative UK provider. The report argues that vendor lock-in should not be seen as inevitable and calls for a strategy to end lock-in across the public sector, diversify suppliers and strengthen digital resilience.

Reliance on a small number of US-based providers represents a clear vulnerability, with ambitions to digitally transform public services potentially “at the mercy” of foreign actors, according to the committee.

To address this, there should be greater clarity over the government’s definition of technological sovereignty, and a clear strategy to deliver it. Such a strategy would help identify where the UK needs sovereign capabilities, and to support the development of domestic alternatives to existing providers in critical sectors – reportedly a priority for the Chancellor.

The report finds that the lack of clarity extends to the heart of the government’s vision for a modern digital state. Whilst welcoming the ambition, the report notes that the government has yet to set out what becoming a “digital state” means in practice. There is no clear plan to translate ambition into reality and and the digital roadmap lacks metrics to assess progress. Challenges have been compounded by disruption associated with the Government Digital Service’s (GDS) move to DSIT, with delivery “hamstrung.”

The report calls for the appointment of a Cabinet-level minister to drive cross-government delivery of digital transformation. Alongside this, the government should undertake an urgent review of GDS’ remit, appoint a permanent secretary-level Government Chief Digital Officer and head of GDS to support this new political figurehead, and publish a detailed delivery plan to ensure accountability.

The report also highlights serious data hygiene deficiencies across the public sector, which can only be overcome through a process of cultural transformation. It calls for a strategy to deliver this, with greater acknowledgement of the fact that public sector bodies hold citizens’ data on trust and therefore must meet the highest standards.

Without addressing the underlying weaknesses in the UK's current digital infrastructure and improving data hygiene, it would be "irresponsible" to roll out a digital ID system, says the committee. Public trust and consent are fundamental to the successful digital transformation of public services - failing to effectively protect citizens' data held or processed for digital ID would have far-reaching consequences.

Chair comment

Dame Chi Onwurah MP, Chair of the Science, Innovation and Technology Committee, said:

"We welcome the government's intentions to make the UK a 'truly digital state,' but it's not clear how this will be delivered. Without a detailed and measurable plan, it risks falling short – but there's still time to put this right. My committee's report sets out how, including by putting clear delivery metrics in place and giving the Government Digital Service the authority and leadership it needs to drive this transformation. Ministers must also prioritise driving up standards of data hygiene across the public sector in order to build trust.

"A critical part of this transformation should include reducing the UK's dependence on a small number of big US tech companies like Palantir. Vendor lock-in isn't inevitable, and the current position leaves us seriously exposed. The UK can and should be aiming for technology sovereignty in critical parts of our public sector and supporting domestic alternatives through smarter procurement.

"Only once the foundations of the UK's digital infrastructure are secure, and public trust has been gained, should the government proceed with its planned digital ID. The success or failure of this project will be a defining test of its wider digital transformation ambitions."

MPs urge Labour to ditch £330m Palantir software contract with NHS

The Guardian · 9 July 2026

https://www.theguardian.com/politics/2026/jul/09/mps-urge-labour-to-ditch-330m-palantir-software-contract-with-nhs



A second parliamentary committee has urged Labour to scrap Palantir's £330m contract with the NHS, increasing pressure on the next prime minister over government deals with the US tech company.

MPs on the health and social care select committee want the NHS to cut ties with Palantir and find a replacement for its system, which is supposed to unify and analyse huge amounts of often highly sensitive NHS health data.

Palantir, cofounded by the Trump-supporting tech billionaire Peter Thiel, also works for the US, Israeli and UK militaries, and the cross-party group of MPs cited "serious mistrust" among the public and the medical profession of its NHS system, contested evidence of its benefits, and cited the availability of other tools that could deliver similar results.

Last month the science and technology committee called for the government to exercise a February 2027 break clause in the deal and either develop an in-house replacement to the federated data platform (FDP) or look for a UK alternative.

Up to 117 NHS data and technology workers broke ranks to call for the deal to be axed, warning that patient privacy protections were inadequate and that a loss of public trust could harm the quality of vital health data.

In a letter to the health secretary, James Murray, they said: "Data completeness will be compromised by the erosion of patient trust; data privacy protections are inadequate and data structures are at risk of misuse."

One signatory, a senior data professional who requested anonymity, said: "The FDP has not shown me any significant technological benefits whatsoever. A frankly mediocre software is being forced on NHS data systems at the expense of patient trust, professional integrity and the fundamental values of the NHS. How can we say we want to use software for saving lives when that same software is also used to kill and ruin lives?"

Alex Karp, Palantir's chief executive, last year responded to a claim that its technology kills Palestinians, by saying: "mostly terrorists, that's true".

In May, Sadiq Khan, the London mayor, blocked the Metropolitan police from awarding a £50m contract to Palantir after saying Londoners wanted to see public money being paid to companies that "share the values of our city". City Hall cited "serious concerns" about how the deal had been struck. Palantir is challenging the decision in the high court.

Layla Moran, the Liberal Democrat chair of the health committee, said: "Little by little, the government's arguments for sticking with the FDP has unravelled. So in the interest of public confidence in the NHS and the security of their medical information, we believe it is time to crack on with preparations to find an alternative."

Palantir has held the contract with NHS England since 2023. But the use of its AI-powered software to support military strikes in Gaza and Iran, US ICE immigration raids and access to NHS data has stoked concern. Palantir's UK and Europe chief, Louis Mosley, has said critics of its NHS role have "chosen ideology over patient safety" and insisted the technology has helped cut surgery waiting lists and accelerate cancer diagnoses.

A spokesperson for Palantir said: "Palantir software is helping to deliver better patient care – including 110,000 additional operations to date, a 15% reduction in discharge delays and a 6.8% increase in patients finding out whether they need cancer treatment within 28 days … But that is what it is – software. How that software is used is controlled by the NHS trusts who use it, with data – legally and contractually – only able to be processed strictly in accordance with their instructions."

The health workers union, Unison, said: "The government cannot allow the NHS to be captured by firms like Palantir which prioritise profits over ethics."

Helga Pile, the union's head of health, said: "The health service needs to retain the ownership and control over patient data systems which are highly sensitive and critical infrastructure. Patients, staff and MPs are calling for Palantir's contract to be discontinued. NHS England must listen to them to restore public confidence."

A spokesperson for the Department of Health and Social Care said: "The government is reviewing this contract ahead of the February 2027 break clause. Thousands more patients are benefiting from the NHS federated data platform every month, with an extra 110,000 of them having undergone procedures in operating theatres, while also reducing the number of unnecessary days patients stay in hospital following treatment by a seventh."

Palantir boss: MPs who want to block us from NHS are ‘noisy minority’

The Telegraph · 29 July 2026

https://www.telegraph.co.uk/news/2026/07/29/palantir-boss-mps-block-nhs-noisy-minority/



MPs who want to block Palantir from the NHS are a "noisy minority", according to the software company's UK boss.

Louis Mosley said those who were critical of the US tech giant were from "a certain end of the political spectrum".

The firm, which has £600m of public sector contracts, including with the NHS and the Ministry of Defence (MoD), has faced growing criticism from the Left regarding its work with US immigration authorities and the Israeli military.

There has also been debate within the health service about whether the software has improved waiting lists, with the Health Foundation think tank disputing claims it had cut the number of patients facing delayed discharge by 15 per cent.

Mr Mosley said it was "up to the British government" to decide whether and how to use Palantir software, as he dismissed concerns about its staff accessing patient data.

'A lot of noise about us'

He told Times Radio: "There is a lot of noise about us. It does come from, I would say, a minority. It's a specific end of the political spectrum. It's a noisy minority. It's people like Zack Polanski.

"They have a big issue with the way in which our software is used thousands of miles away in other countries, specifically America and Israel.

"My view, and by the way, I'm very happy to debate that. I will defend it. But when it comes to the NHS, my view is what matters is, can the NHS control the data when it's using Palantir? Is patient data safe if it's using Palantir? And secondly, are the Palantir tools working? The answer to both those questions is a categoric yes."

He said that no one should be worried about Palantir staff accessing NHS data.

"You could think of Palantir a bit like a plumbing system that the NHS has deployed to manage the flow of data around hospitals and between hospitals. We don't have an economic interest in the data. We're not deriving any value. It's not a business model."

"And it's also against the law. The way in which the law works in this country is it makes a distinction between data controllers and data processors. Data controllers in this case are the NHS. That means it's your data. You own it. We're the processor."

The government's health and social care committee has recommended that the NHS cut short its deal with Palantir.

There is a break clause in the £330m seven-year contract next March, which would need to be exercised by the end of this year.

However, the committee has been criticised for allowing Martin Wrigley MP, an anti-Palantir campaigner, to ask questions in a guest appearance during an evidence session rather than being called as a witness.

The chief executive of Palantir UK also rejected the conclusions of the technology committee, which called the firm's involvement in UK defence an "unacceptable point of weakness".

"You'd have to ask the members of the select committee why they came to that conclusion. I would strongly reject it. And I think the last 10 years of work with the MoD would suggest that it's not correct."

Correcting the Record: Response to the EFF January 15, 2026 Report on Palantir

Palantir Blog · 27 January 2026

https://blog.palantir.com/correcting-the-record-response-to-the-eff-january-15-2026-report-on-palantir-4b3a12536cd2



Editor's Note: This blog post responds to allegations published by the Electronic Frontier Foundation (EFF) in relation to Palantir's work with Immigration and Customs Enforcement (ICE). We believe it's important to address misconceptions (as we have previously) about our technology and business practices with transparency and factual accuracy.

Introduction

The Electronic Frontier Foundation (EFF) has long served as a venerable digital rights organization focused on online civil liberties issues. In the past, EFF's intellectual rigor and careful analysis have served an important function in public discussions about the use of technology. That is why it is particularly important to address instances in which EFF's recent reporting on Palantir got the story wrong.

EFF's January 15, 2026 report, "ICE Using Palantir Tool That Feeds On Medicaid Data" makes a number of misleading and incorrect claims about Palantir's work with the US Federal Government and ICE specifically. Palantir was unfortunately not contacted during the creation of their report, which would have offered us the opportunity to address many of the misconceptions and previously rebutted claims therein.

In the spirit of transparency and open discourse, we outline and respond to these below.

Specific Claims to Address

Here we take specific quotes from the EFF report and, with the aim of good faith transparency, walk through where there may be partial or whole truths, complete inaccuracies, or misleading insinuations that require clarification. We strive to be as exhaustive as possible. We are open to engaging further on these claims with any stakeholders earnestly interested in knowing the real story.

"We also warned about the danger of the Trump administration consolidating all of the government's information into a single searchable, AI-driven interface with help from Palantir, a company that has a shaky-at-best record on privacy and human rights."

- This statement is riddled with tenuous associations and previously rebutted claims. As their core evidence, EFF cites a June 2025 EFF report that uses as a primary reference allegations presented in the May 30, 2025 New York Times article. Palantir thoroughly debunked this NYT article and its core allegations previously: see Correcting the Record: Responses to the May 30, 2025 New York Times Article on Palantir.

- To be absolutely clear, Palantir is not working on any master database project to unify databases across federal agencies. Palantir has not proposed the US Government build a "master list" for the surveillance of citizens, nor have

we been asked to consider building such a system for any customer.

- EFF's statements misrepresent the fundamental reality of government data sharing across agencies, which is subject to, and governed by, data sharing agreements and government oversight. Further, Palantir is not facilitating widespread and unfettered data sharing across our government customers, as each customer instance of our software is legally, technically, and operationally distinct.

- The supposed evidence alleging Palantir's poor "privacy" record cites a September 2020 Privacy International & No Tech For Tyrants report directed at our UK NHS work that reads more as anti-Palantir polemic than factual inquiry. It rehashes a litany of debunked claims about Palantir's supposed predictive policing work, hollow commitment on PCL protections, lack of transparency in contracting, and questionable relationship to the UK Government. Those allegations have all been addressed in detail on multiple occasions, but notably also in the published annex of that very same report. For more on Palantir's privacy and governance safeguards, see 1, 2, 3, 4.

- Similarly, the statement on Palantir's "human rights" record cites to a September 2020 report from Amnesty International, the published annex of which also includes our rebuttals to the claims of the same report. Palantir takes a rigorous approach to respecting human rights, from the development to customer use of our products. We also actively support humanitarian missions around the world, from fighting human trafficking, to disaster relief, to the defense of sovereign nations.

"Palantir is working on a tool for Immigration and Customs Enforcement (ICE) that populates a map with potential deportation targets, brings up a dossier on each person, and provides a "confidence score" on the person's current address," 404 Media reports today. "ICE is using it to find locations where lots of people it might detain could be based."

- EFF's restatement of claims by 404 Media is misleading. The ELITE tool is used for prioritized enforcement to surface the likely addresses of specific individuals, such as those with final orders of removal or with high severity criminal charges. The purpose of this tool is identifying the location of known foreign nationals who meet criteria for removal, not for mass prioritization of "locations where lots of people it might detain could be based."

- Palantir's role in this tooling centers on data integration, enabling ICE to incorporate data sources to which it has access (including DHS data and limited data shared by other agencies under a data sharing agreement permitting it to be used for immigration enforcement purposes). This data includes information needed by ICE officers and agents to accurately assess whether an individual can be removed, for instance whether an individual has a court-ordered stay on their removal, or whether they have status that precludes their removal.

"The tool — dubbed Enhanced Leads Identification & Targeting for Enforcement (ELITE) — receives peoples' addresses from the Department of Health and Human Services (which includes Medicaid) and other sources, 404 Media reports based on court testimony in Oregon by law enforcement agents, among other sources."

- The EFF report is misleading in suggesting that ICE may have, through Palantir software, unfettered access to "multiple government sources including HHS records." While there are instances of limited information sharing between select component agencies of HHS and ICE pursuant to lawful authority, these instances are governed by the same requirements that apply to other components of government under the Privacy Act and other applicable laws and policies.

- The ELITE tool reported in 404 Media is a component of the capabilities provided under a previously reported pilot program with ICE. As the article explains (and is detailed above), the application supports workflows aimed at prioritizing specific immigration enforcement actions.

- Finally, Palantir's platforms are built with an indelible audit log — all platform interactions, from data ingest onwards, are subject to detailed logging. Palantir's platforms are designed with privacy, security, and auditability at their core which, as we've said before, make them exceptionally poor tools for abuse.

"This kind of consolidation of government records provides enormous government power that can be abused. Different government agencies necessarily collect information to provide essential services or collect taxes, but the danger comes when the government begins pooling that data and using it for reasons unrelated to the purpose it was collected."

- Indeed, these are real risks. Which is why we've clarified — in responses to both the NYT and the Wyden/AOC letter — that we are not building a "master database" for the Federal Government. Such a hypothetical project would be fundamentally at odds with Palantir's values and our commitment to work in support of liberal democracies.

- Moreover, we support efforts to enforce strong privacy protections in government systems, which is why we are advocating for Privacy Act reform. Palantir has been on the cutting-edge of privacy-enhancing technology since our company's inception. It is the cornerstone of what we do.

“While couched in the benign language of eliminating government ‘data silos,’ this plan runs roughshod over your privacy and security.”

- This statement mistakenly represents Palantir’s work with the US Government. All of Palantir’s government contracts and the corresponding deployments of software instances are unique to the contracting agency, with legal, procedural, and technical guardrails in place to protect each agency’s data.

- It is also wrong to suggest that any sharing (or so-called ‘merging’) of data across partner agencies implies malicious intent. One need only query terms like ‘government data usage’ or ‘data sharing agreement’ to understand the shape and commonality of efforts by Federal Government agencies to cooperate on initiatives involving limited, purpose-specific information sharing to support legitimate cross-agency workflows. This is a routine aspect of government functioning in the modern age. Further, privacy is not guaranteed by a decentralized, siloed series of government systems in which there is no audit trail or visibility into how data is shared or used. All that is guaranteed by such a fragmented data landscape is that abuse of data sharing and privacy will be harder to systematically guard against or monitor.

“But litigation isn’t enough. People need to keep raising concerns via public discourse and Congress should act immediately to put brakes on this runaway train that threatens to crush the privacy and security of each and every person in America.”

- We couldn’t agree more about the importance of public discourse. But the obligation of engaging in public discourse is to do so honorably, in good faith, and with a commitment to facts over all. We are disappointed by the resurfacing of previously debunked or otherwise incorrect claims furthering a narrative that Palantir is leading an effort to spy on Americans and run roughshod over the privacy of our fellow citizens. This is not happening and the principles that underpin our work guard against such misuse of our products.

Conclusion

Palantir has long shared EFF’s commitment to privacy, civil liberties, and security for Americans, and we would have welcomed the opportunity to address the many incorrect claims in their report prior to publication.

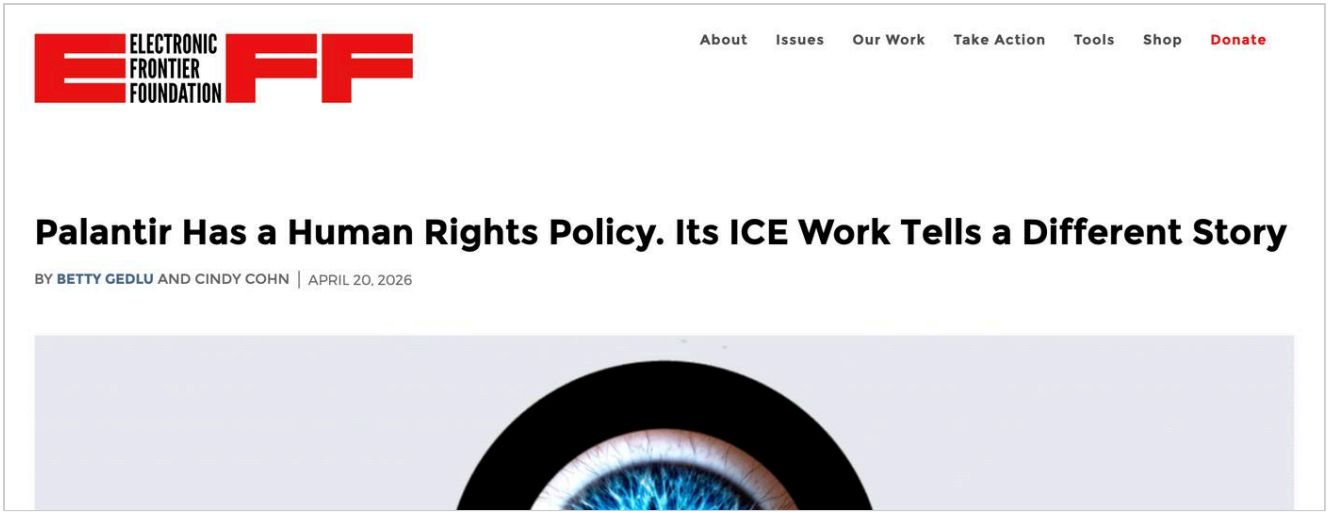
We recognize the complexity that our work with the US Government presents, especially during such polarizing times. Palantir has worked with the US Government for two decades and across multiple administrations, and we see it as our enduring mission to support those who work tirelessly to help our institutions function, even when that work is controversial.

We are committed to transparency around the development and responsible use of our products, and hope that future reporting on us focuses on good faith dialogue — that which prioritizes the preservation of humanity in the global commons over conformity to hysteria and polarization. As Hannah Arendt said, “we humanize what is going on in the world and in ourselves only by speaking of it, and in the course of speaking of it we learn to be human.”

Palantir Has a Human Rights Policy. Its ICE Work Tells a Different Story

Electronic Frontier Foundation · 20 April 2026

https://www.eff.org/deeplinks/2026/04/palantir-has-human-rights-policy-its-ice-work-tells-different-story



For years, EFF has pushed technology companies to make real human rights commitments—and to live up to them. In response to growing evidence that Palantir's tools help power abusive immigration enforcement by ICE, we sent the company a detailed letter asking how the promises in its own human rights framework extends to that work.

This post explains what we asked, how Palantir responded, and why we believe those responses fall short. EFF is not alone in raising alarms about Palantir; immigrants' rights groups, human rights organizations, journalists, and former employees have raised similar concerns based on reports of the company's role in abusive immigration enforcement. We focus here on Palantir's own human rights promises.

At the outset, we appreciate that Palantir was willing to engage respectfully, and we recognize that confidentiality and security obligations can limit what it can say. Nonetheless, measured against Palantir's own human rights commitments, its decision to keep powering ICE with tools used in dragnet raids and discriminatory detentions is indefensible. A good-faith application of those commitments should lead Palantir to end its contract with ICE, and refuse new, or end current, contracts with any other agency whose work predictably violates those commitments.

Palantir's Public Promises

Palantir has long said it performs comprehensive human rights analysis on its work. It has also worked with ICE for years, apparently in a more limited capacity than today. It has publicly embraced the UN Guiding Principles on Business and Human Rights, the Universal Declaration of Human Rights, and the OECD Guidelines for Multinational Enterprises. Additionally, in its response to EFF, Palantir says its legal responsibilities are only "the floor" for broader risk assessments.

That was the point of our letter. We asked what human rights due diligence Palantir conducted when it first contracted with ICE and DHS; whether it performed the "proactive risk scoping" it advertises, how it reviews work over time, what it has done in response to reports of misuse, and whether it has used "every means at [its] disposal"—including contract provisions, third-party oversight, and termination—to prevent or mitigate harms.

For the most part, Palantir did not answer our accountability questions. It did correct one point: Palantir says it does not currently work with CBP, and available evidence supports that, though it also made clear it could work with CBP in the future.

Palantir also raised a red herring it often deploys in response to criticism. It denied building a 'mega' or 'master' database for ICE and denied creating a database of protesters, which some ICE agents have claimed to have been built. We call it a red herring because those denials sidestep the central issues: what capabilities Palantir's tools actually provide to ICE.

To be clear, EFF has never claimed that Palantir is building a single centralized database. Our concern is grounded in how Palantir's tools allow ICE to query and analyze data from multiple databases through a unified interface—which from an agent's perspective can be a distinction without a difference.

In the sections that follow, we compare Palantir’s account of its work for ICE with evidence about how its tools seem to be used, and explain why legality, internal process, and sustained “engagement with the institutions whose vital tasks exist in tension with certain human rights” are no substitute for real human rights due diligence—because respect for human rights must be measured by outcomes, not just process.

Palantir’s ICE Work Undermines Its Own Standards

Palantir says ICE uses its ELITE tool for “prioritized enforcement”: to surface likely addresses of specific people, such as individuals with final orders of removal or high-severity criminal charges. But according to sworn testimony in Oregon, ICE agents use ELITE to determine where to conduct deportation sweeps, and the system “pulled from all kinds of sources” to identify locations for raids aimed at mass detentions, including information from the Department of Health and Human Services such as Medicaid data. A leaked ELITE user guide for 'Special Operations' also instructs operators to disable filters to "display all targets within a Special Operations dataset." Those details directly conflict with Palantir’s narrow description of ELITE’s role.

Additionally, Palantir's response leans on legal authority and the Privacy Act. But it does not identify any specific lawful basis for using Medicaid data in this way or explain how its software enables that access. Even if a legal theory exists, turning sensitive medical information into fuel for dragnet sweeps is hard to reconcile with its commitments to privacy, equity, and the rights of impacted communities. Its own human rights framework requires grappling with foreseeable harms its products may enable, not just invoking possible legal authorization.

Reporting shows that many people detained by ICE had no criminal record, much less a serious one, and in many cases no final order of removal. An overwhelming percentage of those detained were, or appeared to be, from Central and South America, and nearly one in five ICE arrests were street arrests of a Latine person with neither a criminal history nor a removal order.

These facts raise obvious questions about discriminatory impact, racial profiling, and whether Palantir's tools are facilitating detention practices far broader than the company claims. Palantir's response does not meaningfully engage those questions, despite the company's commitments to non-discrimination and due process.

EFF’s letter asked Palantir to explain how it is honoring its commitments to civil liberties in light of reports linking Palantir-owned systems to facial recognition and other tools used to identify and target people engaged in observing and recording law enforcement, including in connection with the deaths of Renée Good and Alex Pretti. The letter also cites an incident in which an officer scanned protesters’ and observers’ faces and threatened to add their biometrics to a “nice little database.” Palantir’s response denies involvement in any such database.

A narrow denial about a single database does not answer the broader question: if ICE, its customer, claims it has this capability, what has Palantir done to ensure its tools are not used to chill protected speech, retaliate against observers, or facilitate targeting of people engaged in First Amendment-protected activity? For a company that claims to value democracy and civil liberties, this is not a marginal issue; it goes to the heart of its human rights commitments.

Legality, Process, and Engagement with ICE Are Not Human Rights Standards

As mentioned above, Palantir leans heavily on legal compliance. It says government data sharing is “subject to, and governed by, data sharing agreements and government oversight” and that any sharing it facilitates is done according to “legal and technical requirements, including those of the Privacy Act of 1974.” It describes its role in ELITE as “data integration,” enabling ICE “to incorporate data sources to which it has access,” including data shared under inter-agency agreements.

EFF is very familiar with the Privacy Act—we are suing the Office of Personnel Management over it currently. But Palantir’s response does not clarify how ICE legally has access to this information, how Palantir ensures that it follows those legal processes, or how Palantir’s software may have enabled access in the first place. More critically, that is still a legal answer to a human rights question, and legal compliance alone is insufficient as a human rights standard.

Human rights due diligence requires assessing foreseeable harms, responding to credible evidence of abuse, and changing course when the facts demand it—something Palantir, on paper, recognizes. That’s why it stresses that its legal responsibilities are only “the floor for [its] broader risk assessments,” pointing to the way it built toward GDPR-style data protection principles and incorporated international humanitarian law principles before those requirements were formalized. If those commitments mean anything, Palantir has to explain how specific practices—like enabling ICE to use Medicaid data in dragnet raids—square with that broader standard.

Palantir also leans heavily on process. It points to a “layered approach” to risk, frameworks that purportedly examine multiple dimensions of privacy and equity, and “indelible” audit logs that track how its tools are used. Audit logs are not sufficient for protecting human rights. There is a long history of authoritarian regimes keeping extensive logs of

their human rights abuses. Those structures can be useful for protecting human rights, but only if they are used to detect harm, trigger reassessment, and lead to changes in design, access, support, or contract enforcement when credible reports of abuse emerge.

That is why we pressed Palantir to spell out clearly what reports of misuse Palantir has received, what changes it made, and on what timeline. Again, instead of offering specific examples, Palantir points back to its internal framework and its willingness to "move towards the hardest problems" as evidence of effective efforts. But human rights are an outcome, not just a process.

Human rights due diligence is not a one-time approval at contract signing; under the UN Guiding Principles, it is supposed to be continuous, with new facts triggering reassessment. Complaints, media reports, leaks, litigation, and sworn testimony are exactly the kinds of events that should prompt review. If Palantir has an account for that work—how often it reviews ICE contracts, who conducts the reviews, what triggers them, and how findings reach the Board — it had every opportunity to describe it. Instead, it offered a generic assurance that it remains committed to human rights without engaging in the specifics. Confidentiality may sometimes limit disclosure, but it is no substitute for accountability.

What Needs to Happen Next

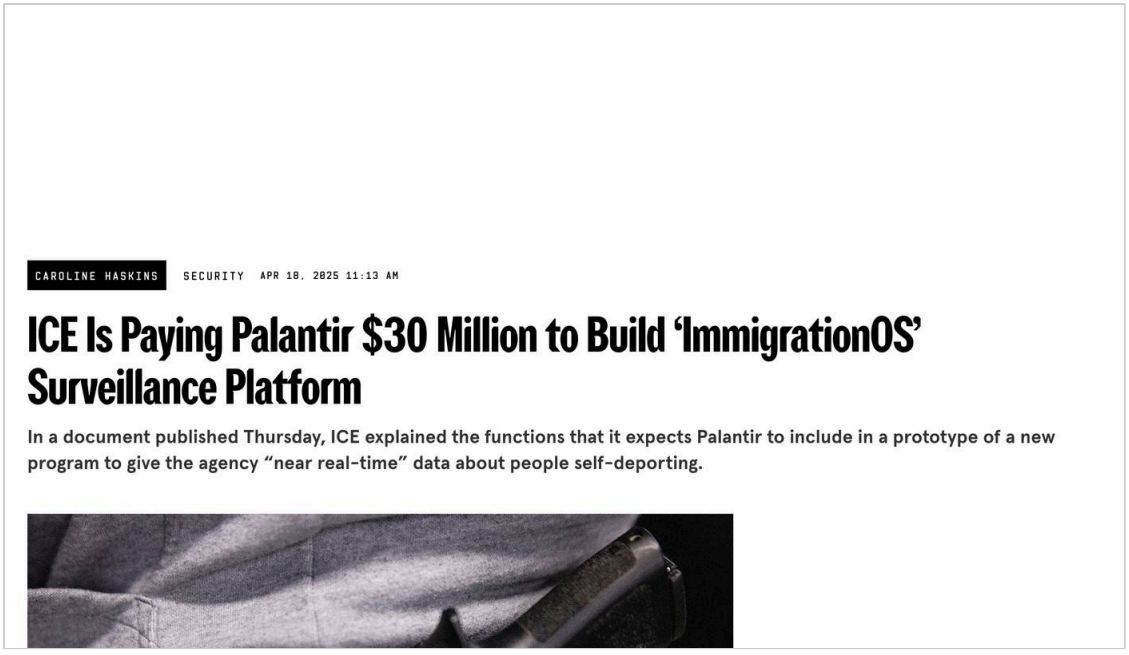
Palantir wants credit for "mov[ing] towards the hardest problems" and engaging with institutions whose missions it says are "in tension with certain human rights" while having a human rights framework. But when the record includes violent raids, dragnet detentions, use of sensitive medical data, discriminatory targeting, retaliation against observers, and deaths tied to immigration enforcement operations, pointing to a values page is not enough; it has to reckon with the results.

Voluntary corporate human rights policies often function as weak accountability mechanisms: companies can tout principles, publish policies, and answer criticism with polished statements while changing very little on the ground. Palantir's response fits that pattern all too well. EFF will continue to challenge its role in abusive immigration enforcement and demanding more accountability for technology vendors whose tools enable human rights violations. We are also happy to continue a dialogue with Palantir to that end. For now, this much is clear: Palantir needs to reconsider its contract with ICE and with all agencies whose work predictably violate human rights.

ICE Is Paying Palantir $30 Million to Build ‘ImmigrationOS’ Surveillance Platform

WIRED · 18 April 2025

https://www.wired.com/story/ice-palantir-immigrationos/



CAROLINE HASKINS SECURITY APR 18, 2025 11:13 AM

ICE Is Paying Palantir $30 Million to Build ‘ImmigrationOS’ Surveillance Platform

In a document published Thursday, ICE explained the functions that it expects Palantir to include in a prototype of a new program to give the agency “near real-time” data about people self-deporting.

Immigration and Customs Enforcement is paying software company Palantir $30 million to provide the agency with “near real-time visibility” on people self-deporting from the United States, according to a contract justification published in a federal register on Thursday. The tool would also help ICE choose who to deport, giving special priority to “visa overstays,” the document shows.

Palantir has been an ICE contractor since 2011, but the document published Thursday indicates that Palantir wants to provide brand-new capabilities to ICE. The agency currently does not have any publicly known tools for tracking self-deportation in near real-time. The agency does have a tool for tracking self-reported deportations, but Thursday’s document, which was first reported by Business Insider, does not say to what degree this new tool may rely on self-reported data. ICE also has “insufficient technology” to detect people overstaying their visas, according to the Department of Homeland Security. This is particularly due to challenges in collecting "biographic and biometric" data from departing travelers, especially if they leave over land, according to Customs and Border Protection.

The agency says in the document that these new capabilities will be under a wholly new platform called the Immigration Lifecycle Operating System, or ImmigrationOS. Palantir is expected to provide a prototype of ImmigrationOS by September 25, 2025, and the contract is scheduled to last at least through September 2027. ICE’s update to the contract comes as the Trump administration is demanding that thousands of immigrants “self-deport,” or leave the US voluntarily.

ICE and Palantir did not respond for comment.

According to the document, ImmigrationOS is intended to have three core functions. Its “Targeting and Enforcement Prioritization” capability would streamline the “selection and apprehension operations of illegal aliens.” People prioritized for removal, ICE says, should be “violent criminals,” gang members, and “visa overstays.”

Its “Self-Deportation Tracking” function would have “near real-time visibility into instances of self-deporation,” the document says. The document does not say what data Palantir would use for such a system, but ICE says it aims to “accurately report metrics of alien departures from the United States.” The agency stipulates that this tool should also integrate with “enforcement prioritization systems to inform policy” but does not elaborate on these systems or policies.

Meanwhile, the “Immigration Lifecycle Process” function would streamline the “identification” of aliens and their “removal” from the United States, with the goal of making "deportation logistics” more efficient.

In a “rationale” section, ICE claims that it has an “urgent and compelling” need for ImmigrationOS’s capabilities. Without them, ICE claims, it would be “severely” limited in its ability to target the gangs MS-13 and Tren de Aragua,

and abide by President Donald Trump's executive order to expedite deportations.

Palantir, ICE claims, is "the only source that can provide the required capabilities and prototype of ImmogrationOS [sic] without causing unacceptable delays." ICE says the company has developed "deep institutional knowledge of the agency's operations over more than a decade of support."

"No other vendor could meet these timeframes of having the infrastructure in place to meet this urgent requirement and deliver a prototype in less than six months," ICE says in the document.

ICE's document does not specify the data sources Palantir would pull from to power ImmigrationOS. However, it says that Palantir could "configure" the case management system that it has provided to ICE since 2014.

Palantir has done work at various other government agencies as early as 2007. Aside from ICE, it has worked with the US Army, Air Force, Navy, Internal Revenue Service, and Federal Bureau of Investigation. As reported by WIRED, Palantir is currently helping Elon Musk's so-called Department of Government Efficiency (DOGE) build a brand-new "mega API" at the IRS that could search for records across all the different databases that the agency maintains.

Last week, 404 Media reported that a recent version of Palantir's case-management system for ICE allows agents to search for people based on "hundreds of different, highly specific categories," including how a person entered the country, their current legal status, and their country of origin. It also includes a person's hair and eye color, whether they have scars or tattoos, and their license-plate reader data, which would provide detailed location data about where that person travels by car.

These functionalities have been mentioned in a government privacy assessment published in 2016, and it's not clear what new information may have been integrated into the case management system over the past four years.

This week's $30 million award is an addition to an existing Palantir contract penned in 2022, originally worth about $17 million, for work on ICE's case management system. The agency has increased the value of the contract five times prior to this month; the largest was a $19 million increase in September 2023.

The contract's ImmigrationOS update was first documented on April 11 in a government-run database tracking federal spending. The entry had a 248-character description of the change. The five-page document ICE published Thursday, meanwhile, has a more detailed description of Palantir's expected services for the agency.

The contract update comes as the Trump administration deputizes ICE and other government agencies to drastically escalate the tactics and scale of deportations from the US. In recent weeks, immigration authorities have arrested and detained people with student visas and green cards, and deported at least 238 people to a brutal megaprison in El Salvador, some of whom have not been able to speak with a lawyer or have due process.

As part of its efforts to push people to self-deport, DHS in late March revoked the temporary parole of more than half a million people and demanded that they self-deport in about a month, despite having been granted authorization to live in the US after fleeing dangerous or unstable situations in Cuba, Haiti, Nicaragua, and Venezuela under the so-called "CHNV parole programs."

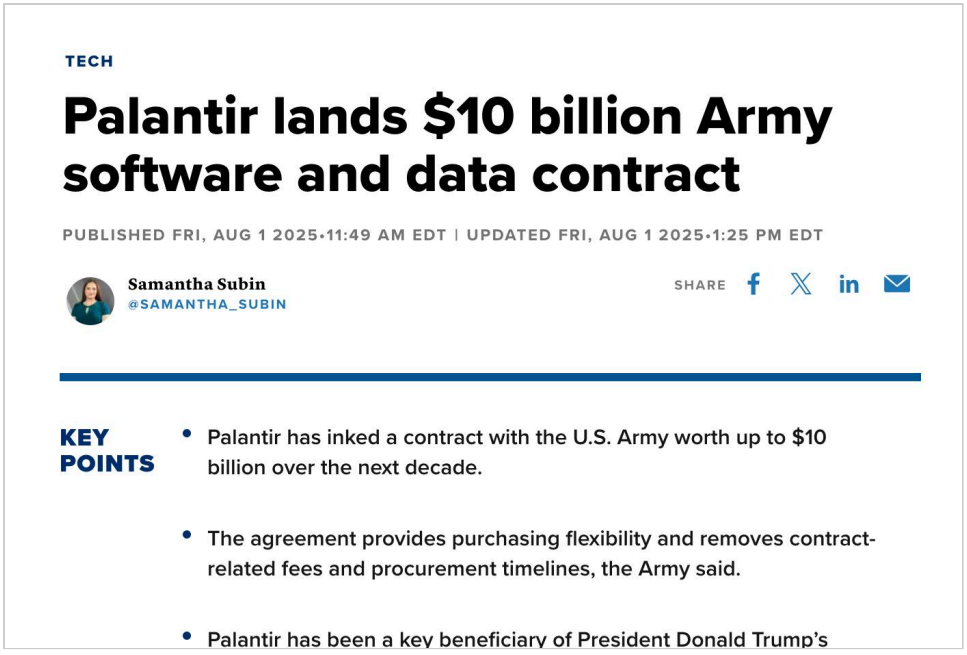
Last week, the Social Security Administration listed more than 6,000 of these people as dead, a tactic meant to end their financial lives. DHS, meanwhile, sent emails to an unknown number of people declaring that their parole had been revoked and demanding that they self-deport. Several US citizens, including immigration attorneys, received the email.

On Monday, a federal judge temporarily blocked the Trump administration's move to revoke people's authorization to live in the US under the CHNV programs. White House spokesperson Karoline Leavitt called the judge's ruling "rogue."

Palantir lands $10 billion Army software and data contract

CNBC · 1 August 2025

https://www.cnbc.com/2025/08/01/palantir-lands-10-billion-army-software-and-data-contract.html



TECH

Palantir lands $10 billion Army software and data contract

PUBLISHED FRI, AUG 1 2025•11:49 AM EDT | UPDATED FRI, AUG 1 2025•1:25 PM EDT

Samantha Subin
@SAMANTHA_SUBIN

SHARE

KEY POINTS

- Palantir has inked a contract with the U.S. Army worth up to $10 billion over the next decade.
- The agreement provides purchasing flexibility and removes contract-related fees and procurement timelines, the Army said.
- Palantir has been a key beneficiary of President Donald Trump's

Palantir has inked a contract with the U.S. Army worth up to $10 billion to meet growing warfare demands over the next decade.

As part of the deal, Palantir will help the military streamline efficiencies while preparing for threats, consolidating 75 total contracts into one enterprise deal, the release states.

The agreement creates a “comprehensive framework for the Army’s future software and data needs” that provides the government with purchasing flexibility and removes contract-related fees and procurement timelines, according to a release.

The deal further cements the company’s role in the U.S. government’s clampdown on cost efficiencies by using artificial intelligence tools in President Donald Trump’s administration. Trump’s Department of Government Efficiency has cut jobs and programs in an effort to curb spending.

Palantir co-founder and CEO Alex Karp has been a vocal proponent of protecting U.S. interests and joining forces on AI to fend off adversaries.

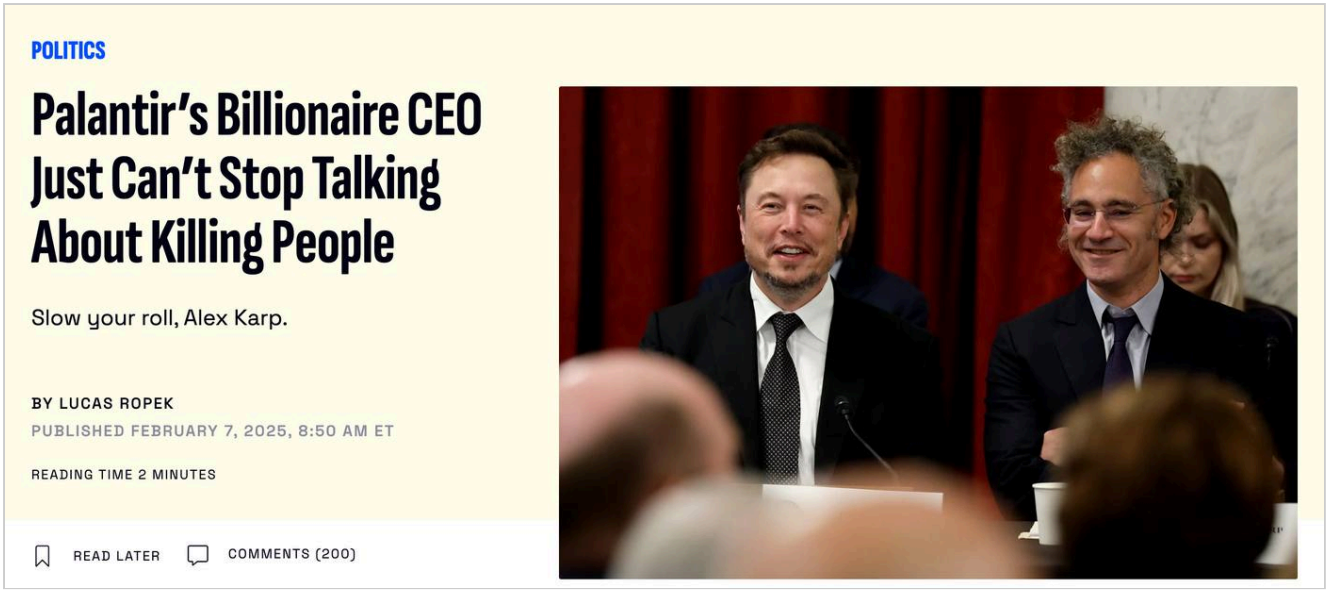
Shares of the Denver-based artificial intelligence software company have more than doubled year to date.

Earlier this year, Palantir delivered its first two AI-powered systems in its $178 million contract with the U.S. Army. In May, the Department of Defense boosted its Maven Smart Systems contract to beef up AI capabilities by $795 million.

Palantir’s Billionaire CEO Just Can’t Stop Talking About Killing People

Gizmodo · 7 February 2025

https://gizmodo.com/palantirs-billionaire-ceo-just-cant-stop-talking-about-killing-people-2000560597



Alex Karp, the creepy CEO of creepy defense contractor Palantir, just can’t stop talking about killing people. During a recent call with investors, the billionaire let it slip that he doesn’t mind a little bloodshed, just so long as the money keeps pouring in.

“Palantir is here to disrupt and make the institutions we partner with the very best in the world and, when it’s necessary, to scare enemies and on occasion kill them,” Karp said, with a smile on his face. The CEO added that he was very proud of the work his firm is doing and that he felt it was good for America. “I’m very happy to have you along for the journey,” he said. “We are crushing it. We are dedicating our company to the service of the West, and the United States of America, and we’re super-proud of the role we play, especially in places we can’t talk about.”

Mother Jones reports that Trump’s re-election and subsequent enactment of police-state style policies have been mighty good for Palantir’s stock— which makes sense since Palantir is a police-state kind of company.

During the same call, Karp mentioned killing people once again when the conversation turned to the fiasco currently unfolding in the U.S. government (said fiasco involves Karp’s fellow billionaire Elon Musk using his Department of Government Efficiency, or DOGE, to dismantle federal agencies, purge the civil service workforce, and generally destroy the functioning of American bureaucracy). “We love disruption, and whatever’s good for America will be good for Americans and very good for Palantir,” Karp said, apparently excited about Musk’s effort. “Disruption, at the end of the day, exposes things that aren’t working,” he continued. “There will be ups and downs. There’s a revolution. Some people are going to get their heads cut off. We’re expecting to see really unexpected things and to win.”

The tech industry’s love of the buzzword “disruption” is humorous and confounding. After all, not all disruption is good disruption. It would be “disruptive” to Karp’s day if, for instance, he was attacked by a grizzly bear, or tripped down his staircase, or got food poisoning, but that doesn’t necessarily mean it would be a good use of his time or beneficial to his overall well-being. Similarly, “disrupting” our nation’s government and allowing people’s heads to be cut off (if only metaphorically) may not actually be good for America.

It’s weird that Karp is so supportive of publicly flogging the U.S. government, since Palantir, itself, was originally seeded with money from that government, and continues to make most of its money from it. Of course, Karp is so rich now it’s possible he just doesn’t see the connection between a functioning bureaucracy and his own well-being. Bloomberg’s billionaire index claims he is worth some $9 billion. With that kind of money, you can afford to kick back and presume that everything’s going to turn out all right in the end.

Thiel's Palantir, Israel Agree Strategic Partnership for Battle Tech

Bloomberg (reprint) · 12 January 2024

https://www.palantir.com/assets/xrfr7uokpv1b/3MuEeA8MLbLDAyxixTsile/9e4a11a7fb058554a8a1e3cd83e31c09/C134184_finaleprint.pdf

Bloomberg

January 12, 2024 | bloomberg.com

Technology

Thiel's Palantir, Israel Agree Strategic Partnership for Battle Tech

- US company will provide services for 'war-related missions'
- Agreement follows co-founders' visit to Israel this week

by Marissa Newman
January 12, 2024

Palantir Technologies Inc., the data analysis firm that provides militaries with artificial intelligence models, has agreed to a strategic partnership with the Israeli Defense Ministry to supply technology to help the country's war effort.

The agreement followed a Thursday meeting between Israeli defense officials and Palantir co-founders Peter Thiel and Alex Karp in Tel Aviv, according to Executive Vice President Josh Harris, who was also in attendance.

"Both parties have mutually agreed to harness Palantir's advanced technology in support of war-related missions," Harris told Bloomberg on Friday. "This strategic partnership aims to significantly aid the Israeli Ministry of Defense in addressing the current situation in Israel."

No further details on the arrangement were disclosed, including what technology would be provided. Palantir last year introduced its AIP, or Artificial Intelligence Platform, an intelligence and decision-making system that can analyze enemy targets and propose battle plans.



Palantir CEO Alex Karp, center, with the head of Israel's Directorate of Defense Research & Development Daniel Gold, center right, and Palantir co-founder Peter Thiel, second right. *Source: Palantir Technologies Inc.*

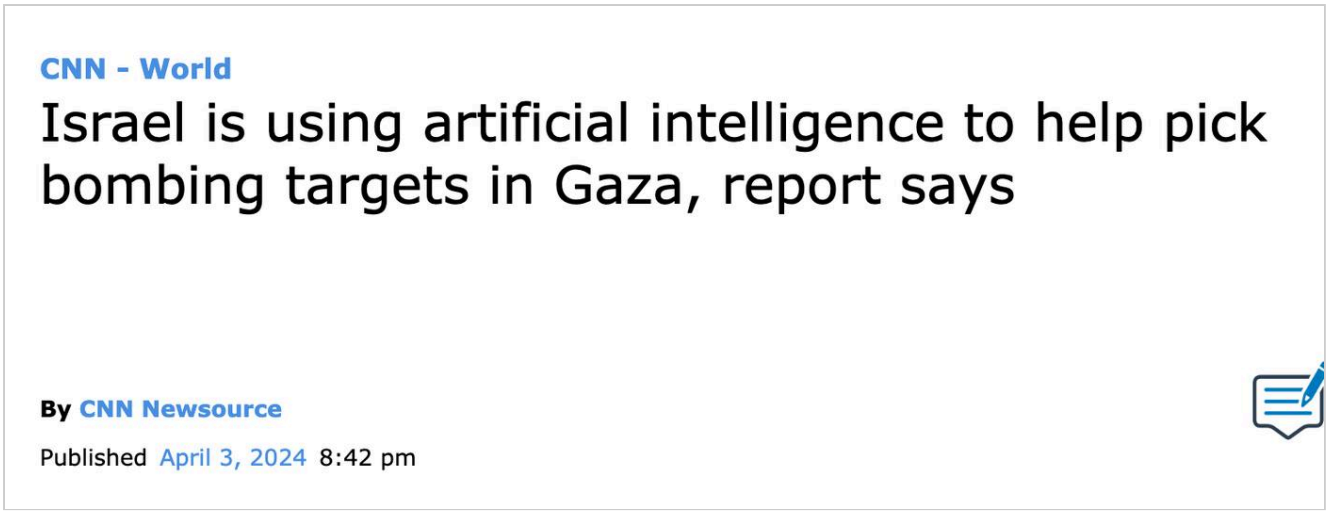
Karp in an interview with Bloomberg on Wednesday said the Denver-based company's products were in great demand in Israel since its war with Hamas began on Oct. 7. Palantir held a board meeting in Tel Aviv for the first time this week in solidarity with Israel.

Posted from *Bloomberg.com*, January 12, 2024, copyright by Bloomberg L.P. with all rights reserved
This reprint implies no endorsement, either tacit or expressed, of any company, product, service or investment opportunity.
#C134184 Managed by The YGS Group, 800.290.5460. For more information visit www.theYGSgroup.com.

Israel is using artificial intelligence to help pick bombing targets in Gaza, report says

CNN (via KION) · 3 April 2024

https://kioncentralcoast.com/news/national-world/cnn-world/2024/04/03/israel-is-using-artificial-intelligence-to-help-pick-bombing-targets-in-gaza-report-says/



By Tara John, CNN

(CNN) — The Israeli military has been using artificial intelligence to help identify bombing targets in Gaza, according to an investigation by +972 Magazine and Local Call, citing six Israeli intelligence officials involved in the alleged program – who also allege that human review of the suggested targets was cursory at best.

The officials, quoted in an extensive investigation by the online publication jointly run by Palestinians and Israelis, said that the AI-based tool was called “Lavender” and was known to have a 10% error rate.

When asked about +972 Magazine’s report, Israel Defence Forces (IDF) did not dispute the existence of the tool but denied AI was being used to identify suspected terrorists. But in a lengthy statement it emphasized that “information systems are merely tools for analysts in the target identification process,” and that Israel tries to “reduce harm to civilians to the extent feasible in the operational circumstances ruling at the time of the strike.”

The IDF said “analysts must conduct independent examinations, in which they verify that the identified targets meet the relevant definitions in accordance with international law and additional restrictions stipulated in the IDF directives.”

However, one official told +972 “that human personnel often served only as a “rubber stamp” for the machine’s decisions” and typically devoted only around 20 seconds to each target – ensuring they are male – before authorizing a bombing.

The investigation comes amid intensifying international scrutiny of Israel’s military campaign, after targeted air strikes killed several foreign aid workers delivering food in the Palestinian enclave. Israel’s siege of Gaza has killed more than 32,916 people, according to the Gaza Ministry of Health, and has led to a spiraling humanitarian crisis where nearly three-quarters of the population in northern Gaza are suffering from catastrophic levels of hunger, according to a United Nations-backed report.

The investigation’s author, Yuval Abraham, previously told CNN in January of his work looking into how the Israeli military has been ”heavily relying on artificial intelligence to generate targets for such assassinations with very little human supervision.”

The Israeli military “does not use an artificial intelligence system that identifies terrorist operatives or tries to predict whether a person is a terrorist,” the IDF statement on Wednesday said. But its analysts use a “database whose purpose is to cross-reference intelligence sources, in order to produce up-to-date layers of information on the military operatives of terrorist organizations.”

Human officers are then responsible for verifying “that the identified targets meet the relevant definitions in accordance with international law and additional restrictions stipulated in the IDF directives,” according to the IDF statement, a process also described by +972.

Night attacks

The magazine also reported that the Israeli army “systematically attacked” targets in their homes, usually at night when entire families were present.

“The result, as the sources testified, is that thousands of Palestinians — most of them women and children or people who were not involved in the fighting — were wiped out by Israeli airstrikes, especially during the first weeks of the war, because of the AI program’s decisions,” it wrote.

The report, citing sources, said that when alleged junior militants were targeted, “the army preferred” to use so-called dumb bombs – unguided missiles which can cause large-scale damage.

CNN reported in December that nearly half of the 29,000 air-to-surface munitions dropped on Gaza last fall were dumb bombs, which can pose a greater threat to civilians, especially in densely populated territories like Gaza.

According to the IDF statement, it does not carry out strikes where the expected collateral damage is “excessive in relation to the military advantage” and makes efforts to “reduce harm to civilians to the extent feasible in the operational circumstances.”

It added that the “IDF reviews targets before strikes and chooses the proper munition in accordance with operational and humanitarian considerations, taking into account an assessment of the relevant structural and geographical features of the target, the target’s environment, possible effects on nearby civilians, critical infrastructure in the vicinity, and more.”

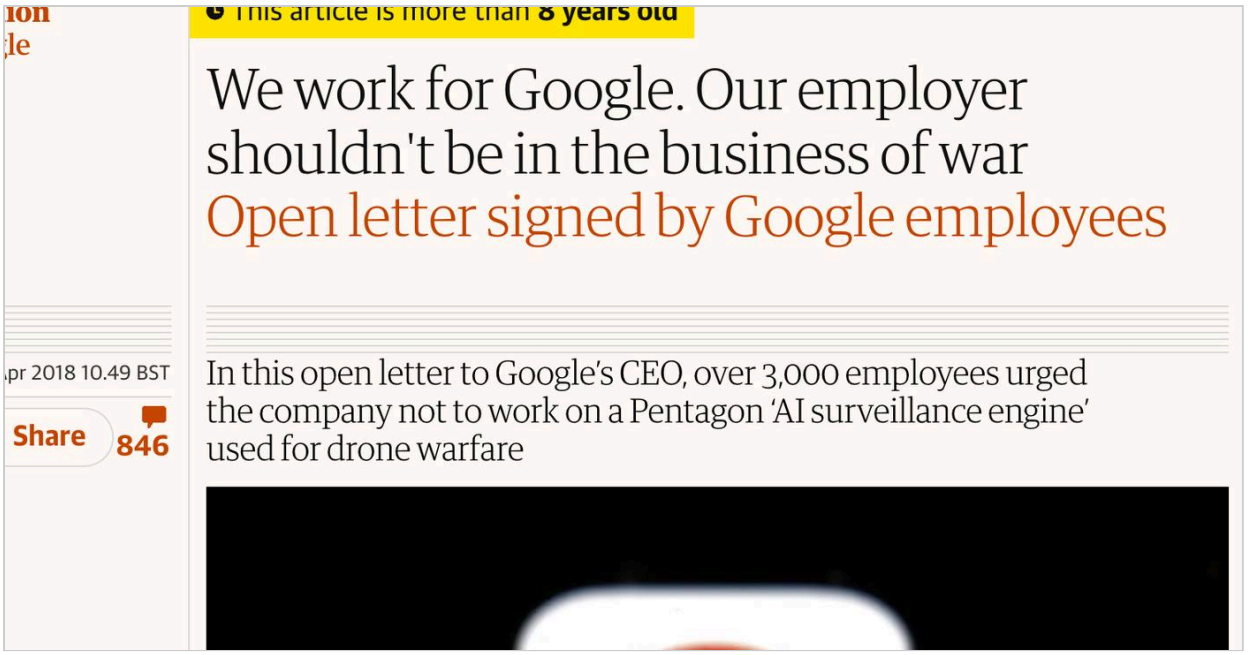
Israeli officials have long argued that heavy munitions are necessary to eliminate Hamas, whose fighters killed more than 1,200 people in Israel and took hundreds of hostages on October 7, sparking the ongoing war.

CNN’s Vasco Cotovio and Mohammed Tawfeeq contributed to this report.

We work for Google. Our employer shouldn't be in the business of war

The Guardian (open letter signed by Google employees) · 4 April 2018

https://www.theguardian.com/commentisfree/2018/apr/04/google-ceo-drones-ai-war-surveillance



Dear Sundar,

We believe that Google should not be in the business of war. Therefore we ask that Project Maven be cancelled and that Google draft, publicize and enforce a clear policy stating that neither Google nor its contractors will ever build warfare technology.

Google is implementing Project Maven, a customized AI surveillance engine that uses "wide area motion imagery" data captured by US government drones to detect vehicles and other objects, track their motions and provide results to the Department of Defense.

Recently, Googlers voiced concerns about Maven internally. Diane Greene responded, assuring them that the technology will not "operate or fly drones" and "will not be used to launch weapons". While this eliminates a narrow set of direct applications, the technology is being built for the military, and once it's delivered it could easily be used to assist in these tasks. This plan will irreparably damage Google's brand and its ability to compete for talent.

Amid growing fears of biased and weaponized AI, Google is already struggling to keep the public's trust. By entering into this contract, Google will join the ranks of companies like Palantir, Raytheon and General Dynamics. The argument that other firms, like Microsoft and Amazon, are also participating doesn't make this any less risky for Google. Google's unique history, its motto "don't be evil", and its direct reach into the lives of billions of users set it apart.

We cannot outsource the moral responsibility of our technologies to third parties. Google's stated values make this clear: every one of our users is trusting us. Never jeopardize that. Ever. This contract puts Google's reputation at risk and stands in direct opposition to our core values. Building this technology to assist the US government in military surveillance – and potentially lethal outcomes – is not acceptable.

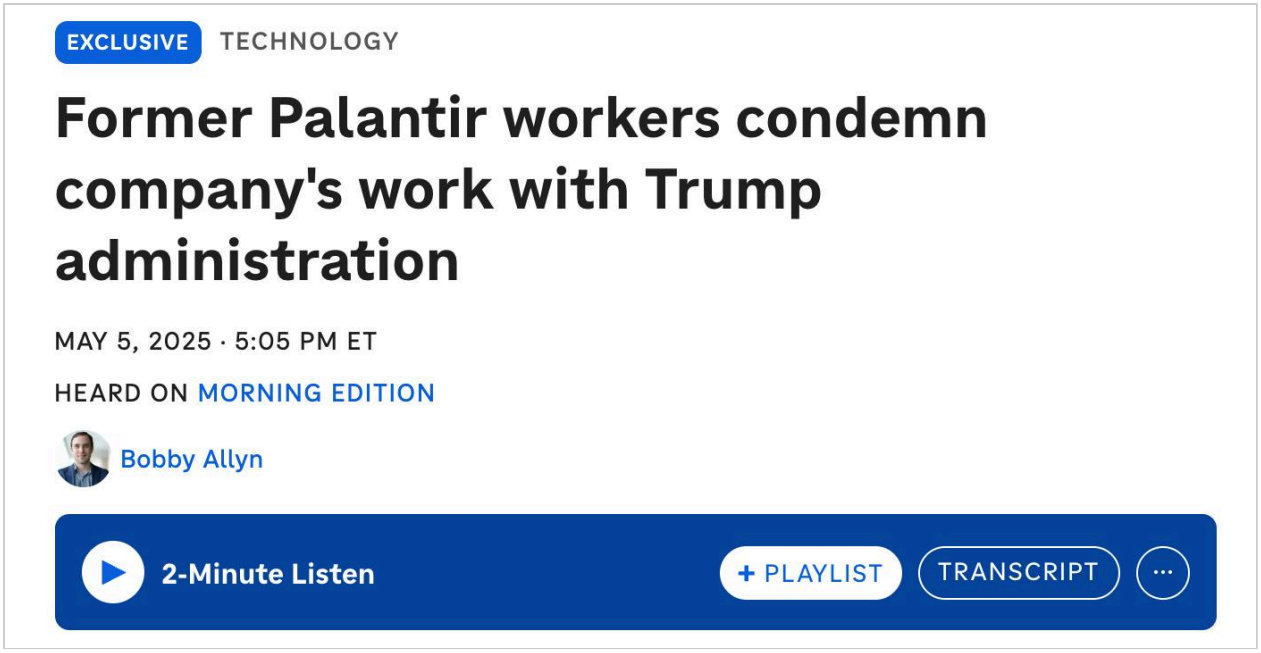
Recognizing Google's moral and ethical responsibility, and the threat to Google's reputation, we request that you: 1. Cancel this project immediately. 2. Draft, publicize and enforce a clear policy stating that neither Google nor its contractors will ever build warfare technology.

- This open letter was originally published in the New York Times

Former Palantir workers condemn company’s work with Trump administration

NPR · 5 May 2025

https://www.npr.org/2025/05/05/nx-s1-5387514/palantir-workers-letter-trump



Thirteen former employees of influential data-mining firm Palantir are condemning the company's work with the Trump administration weeks after Immigration and Customs Enforcement reached a deal to pay Palantir $30 million to provide the agency with "near real-time visibility" into the movement of migrants in the U.S.

In a letter shared exclusively with NPR, the ex-Palantir workers, former software engineers, managers and an employee who worked with the firm's privacy and civil liberties team, say when they joined the powerful tech company, they believed in its code of conduct stating that its software should protect the vulnerable and ensure the responsible development of artificial intelligence.

"Early Palantirians understood the ethical weight of building these technologies," the thirteen former employees wrote in the letter. "These principles have now been violated, and are rapidly being dismantled at Palantir Technologies and across Silicon Valley."

Palantir and the White House did not return requests for comment.

Palantir, co-founded by Trump ally billionaire Peter Thiel, offers data-analyzing software that uses AI to pull information from a multitude of sources and compiles it into charts, tables and heat maps. Its ability to disentangle complex datasets has made its platforms popular with law enforcement and the military.

Palantir's customers include the Israel Defense Forces, the U.S. Department of Defense and dozens of large corporations. Now, its surveillance tools are being deployed by the Trump administration to help speed up the president's goal of deporting 1 million migrants this year. Since Trump was elected, Palantir's share price has surged more than 200%.

Although about a dozen former workers are just a fraction of a company employing 4,000, the letter is still significant. Much of Palantir's work is secretive and few former workers publicly criticize the company after leaving due to non-disparagement agreements most are asked to sign upon leaving. Many former employees also hold company stock. Palantir's market valuation stands at nearly $300 billion — about the same as Bank of America.

While top brass at Silicon Valley's largest companies have been acquiescing to the Trump administration, studies have found that rank-and-file tech workers tend to oppose Trump and hold liberal views on a host of social issues.

Big Tech employees resisted the first Trump administration's policies, including the attempts to restrict immigration and travel from foreign countries, yet the former Palantir employees say the resistance has been muted among tech workers in Trump's second term.

They wrote: "democracy faces escalating threats: biometric data collection on immigrant children, journalists being targeted, science programs defunded, and key U.S. allies, like Ukraine, sidelined. Trump's administration has sought

to greatly expand executive powers while alluding to monarchy," adding that: "Big Tech, including Palantir, is increasingly complicit, normalizing authoritarianism under the guise of a 'revolution' led by oligarchs. We must resist this trend."

The signees highlight what they say is Palantir's "increasingly violent rhetoric," a nod to Palantir chief executive Alex Karp, who publicly boasts about how the company's tools are used to kill enemies. He once jokingly said that Wall Street analysts who "tried to screw" the company should be sprayed with "light fentanyl-laced urine."

The letter also zeros in on Elon Musk's Department of Government Efficiency team, which has hired a number of former Palantir employees to assist in slashing and burning the federal government.

The tech workers admonish the White House for its moves with DOGE and beyond to undo diversity, equity and inclusion initiatives, arguing that those principles "remain essential for critiquing power" and ensuring advanced AI is deployed ethically.

"As Musk's DOGE operation dismantles U.S. government institutions under the guise of exposing corruption, opposition remains silent," the workers write. "Government databases are already erasing references to transgender people and gender-affirming care. These injustices could be facilitated by the very software infrastructure we help build."

The ex-Palantirians write that they are issuing a warning in hopes of triggering a "domino effect" in Silicon Valley.

They say that more tech workers should resist what they see as the misuse of AI and other tools in the Trump administration's actions, including its immigration enforcement and deportation policies that courts and critics say have at times ignored due process rights.

Their aim, they write is "to speak out while we still can, and to work against the dangerous path in the history of technology we are currently heading down towards."

Other Links

Ukraine is using Palantir’s software for ‘targeting,’ CEO says
Reuters · 2 February 2023
https://www.reuters.com/technology/ukraine-is-using-palantirs-software-targeting-ceo-says-2023-02-02/

‘Lavender’: The AI machine directing Israel’s bombing spree in Gaza
+972 Magazine and Local Call · 3 April 2024
https://www.972mag.com/lavender-ai-israeli-army-gaza/

Philosophy

Killer Robots

Robert Sparrow, Journal of Applied Philosophy 24(1): 62–77 · 2007
https://robsparrow.com/wp-content/uploads/Killer-robots.pdf

Journal of Applied Philosophy, Vol. 24, No. 1, 2007

Killer Robots

ROBERT SPARROW

ABSTRACT *The United States Army's Future Combat Systems Project, which aims to manufacture a 'robot army' to be ready for deployment by 2012, is only the latest and most dramatic example of military interest in the use of artificially intelligent systems in modern warfare. This paper considers the ethics of the decision to send artificially intelligent robots into war, by asking who we should hold responsible when an autonomous weapon system is involved in an atrocity of the sort that would normally be described as a war crime. A number of possible loci of responsibility for robot war crimes are canvassed: the persons who designed or programmed the system, the commanding officer who ordered its use, the machine itself. I argue that in fact none of these are ultimately satisfactory. Yet it is a necessary condition for fighting a just war, under the principle of* jus in bellum, *that someone can be justly held responsible for deaths that occur in the course of the war. As this condition cannot be met in relation to deaths caused by an autonomous weapon system it would therefore be unethical to deploy such systems in warfare.*

Introduction

The United States Army's Future Combat Systems Project, which aims to manufacture a 'robot army', to be ready for deployment by 2012, is only the latest and most dramatic example of the military's interest in the use of artificially intelligent systems in modern warfare. A number of 'autonomous weapon systems' are already in use in armed forces around the world, including cruise missiles, torpedoes, submersibles, robots for urban reconnaissance, Uninhabited Aerial Vehicles and Uninhabited Combat Aerial Vehicles; more are in production or on the drawing board. The military remains one of the largest sources of funding for research into robotics and artificial intelligence technology. As these technologies improve it seems inevitable that they will be used more and more in warfare. This paper considers the ethics of a decision to send artificially intelligent robots into war, by asking who we should hold responsible when an autonomous weapon system is involved in an atrocity of the sort that would normally be described as a war crime.

The Real World of Robots at War

The idea of killer robots may seem somewhat farfetched — something out of science fiction. In fact research into military robotics and military applications of artificial intelligence is already well advanced.[1] I shall devote the next few paragraphs to describing autonomous weapon systems (AWS) that are already in production and that those

© Society for Applied Philosophy, 2007, Blackwell Publishing, 9600 Garsington Road, Oxford, OX4 2DQ, UK and 350 Main Street, Malden, MA 02148, USA.

that are, according to reputable sources, likely to be deployed in the next decade, in order to show that the issues discussed here are more pressing than is generally thought.

Existing autonomous weapons systems include cruise missiles, torpedoes, submersibles, robots for urban reconnaissance, Uninhabited Aerial Vehicles (UAVs) and Uninhabited Combat Aerial Vehicles (UCAVs).

Cruise missiles and torpedoes have for a long time exercised a (very limited) degree of autonomy in determining their approach to their target and this is steadily increasing. The Boeing Corporation's SLAM-ER cruise missile, for instance, now has an Automatic Target Recognition capability that allows it to choose a target once it reaches its area of operations, whereas the Northrop Grumman BAT, which homes in on enemy tanks using their acoustic signature, is described by one source as 'a fully autonomous weapons system that can locate, identify, attack and destroy armoured vehicles'.[2] Of particular note in this context is the US Air Force's Low Cost Autonomous Attack System (LOCAAS). LOCAAS is a turbine powered 'stand off' munition, which is designed to 'autonomously search for, detect, identify, attack and destroy theatre missile defence, surface to air missile systems, and interdiction/armour targets of military interest'.[3] It will be equipped with a Laser Radar system and an Autonomous Target Recognition capability that will allow it to search for and identify targets within a 33 sq. mile area. LOCAAS will also carry a warhead that is capable of three different configurations for use against different types of targets, with the machine itself choosing which to employ in its attack.[4] The perceived success of 'smart weapons' in military conflicts in Iraq, Kosovo and Afghanistan means that they are likely to play a large role in any future wars.[5] Improvements in computing power, component miniaturization, and software engineering mean that they are also likely to keep getting 'smarter'.

Robot surveillance drones/UAVs, reconnaissance robots, and submersibles demonstrate more autonomy, as they must be capable of a wider variety of tasks in more challenging environments. The US Navy is developing an Unmanned Underwater Vehicle (UUV), called the Long-Term Mine Reconnaissance System, which can be launched from a submarine whereupon it can seek out and map enemy beach defences and minefields. The Navy is also researching a much more ambitious UUV called MANTA, which will be capable of fully autonomous operation in order to seek out, attack and destroy enemy submarines.[6]

The majority of military robots in existence today however are UAVs. These high tech descendants of Remotely Piloted Vehicles differ from RPVs in having the ability to take off, fly to their objective, and return, with minimal, or even no, human control. UAVs play an increasingly important role in modern air forces. For example, the uninhabited Global Hawk spy plane is now a vital part of the US Air Forces global surveillance capacity. Unpiloted aircraft were used extensively in NATO military operations in Kosovo, both for purposes of surveillance and target designation. By its own reckoning the US Department of Defense has invested over $3 billion (US) in UAVs in the decade 1990–2000 and will spend some $4 billion in the coming decade.[7]

The launching of air-to-ground missiles from armed 'Predator' robot drones in the course of the US invasion of Afghanistan, received extensive publicity. Although a modified version of a pilotless surveillance plane, the Predator may now be classed as a Uninhabited Combat Aerial Vehicle. UCAVs are designed to attack enemies on the ground and in the air, and are widely predicted to be the future of air power.[8] The use

© Society for Applied Philosophy, 2007

of Predator drones, as well as other UAVs, has become a routine part of combat operations in the current conflict in Iraq.[9]

A recent survey of UCAVs cites four as already in production and another seven under development, the most impressive of which is Boeing's X-45, a futuristic unmanned fighter aircraft.[10] Military interest in UCAVs stems from two main sources. Firstly, uninhabited systems offer the prospect of achieving military objectives without risking the politically unacceptable cost of friendly casualties. Secondly, they are expected to be substantially cheaper than the piloted systems they are intended to replace.[11]

Land warfare involves a more challenging terrain and environment for autonomous systems and so the military has been slower to adopt robots designed for ground combat. Nevertheless, the US Army employs a number of reconnaissance robots for use in urban combat (as does the Israeli army) and is developing more.[12] The US army has recently announced its intention to deploy an armed version of a bomb disposal robot, designed to serve as a remotely operated weapons platform, in Iraq.[13] Both UAVs and UCAVs have an obvious role in ground support; the US Marine Corps Warfighting Laboratory is testing a backpack portable UAV called Dragon Eye and a larger, more capable, rotor winged UAV called Dragon Warrior, for use by their infantry.[14]

However, the most ambitious envisioned deployment of AWS is actually intended for land warfare — the United States Army's Future Combat Systems (FCS) project. Developed in collaboration with the Defence Advanced Research Projects Agency, the FCS program has as its main objective the replacement of the US Army's main battle tank with a system (or 'system of systems') that will be capable of rapid deployment to any location in the world in the hold of a C130 transport plane and of integrating the most advanced battlefield technologies currently available, as well as those likely to become available in the next two decades. It is envisioned that many, perhaps all, components of this system will be capable of unmanned operation. Both manned and unmanned components are expected to operate in conjunction with battlefield surveillance provided by robot reconnaissance units, both in the air and on the ground, and fire support provided by robot aircraft and robot artillery platforms.[15] The emphasis on the role of robotics in the FCS concept is so heavy that media descriptions of the project as involving the creation of a 'robot army' do not seem unreasonable.[16]

Most existing versions of these technologies still require some level of human supervision.[17] However the need for such supervision is steadily decreasing as computing and sensor technology improves. It appears increasingly likely that robots will eventually be entrusted with decisions about target identification and destruction.[18]

Beyond these weapons, which are already on the drawing board, lies the prospect of weapons systems that possess genuine 'artificial intelligence'. According to a number of writers in the field, before the end of the century — and according to some, well before this — machines will be conscious, intelligent, entities with capacities exceeding our own.[19] Given that the military constitute a major source of funding for research into artificial intelligence, it seems inevitable that such AIs will be put to work in military contexts.

How 'Autonomous' are Autonomous Weapon Systems?

It is one thing to point out that these weapons exist, and that more are being developed, it is another to claim that they raise any new ethical issues. The idea that they

© Society for Applied Philosophy, 2007

might derives from taking the claim that at some point in the future there might be *autonomous* weapons systems seriously.

Sometimes the claim that a weapon is 'autonomous' means only that it is capable of acting independently of immediate human control; typically it means that it is a 'fire and forget' system capable of determining its trajectory or pursuing its target to some limited extent. Many existing autonomous weapon systems are of this type. Weapons of this nature do not in themselves raise any ethical questions beyond those raised by other modern long range weapons.

However, much more than that is being claimed for the next generation of intelligent robots. According to a number of writers the robots of the future will be capable of acting on their own, in some more robust sense.[20] Artificially intelligent weapon systems will thus be capable of making their own decisions, for instance, about their target, or their approach to their target, and of doing so in an 'intelligent' fashion.[21] While they will be programmed to make decisions according to certain rules, in important circumstances their actions will not be predictable. However this is not to say that they will be random either. Mere randomness provides no support for a claim to autonomy. Instead the actions of these machines will be based on reasons, but these reasons will be responsive to the internal states — 'desires', 'beliefs' and 'values' — of the system itself. Moreover, these systems will have significant capacity to form and revise these beliefs themselves. They will even have the ability to learn from experience.[22] In practice, this is likely to mean that the actions of these machines will quickly become somewhat unpredictable. As we shall see below, this is in itself enough to raise difficult ethical questions about the use of these weapons.

How far the autonomy of these systems extends beyond this is not clear. In particular, it is unclear at what point a weapon system's autonomy means that *it* is, in some real sense, making the decisions with which it is entrusted. An important reason for the lack of clarity about this question is that the nature and extent of autonomy, even in people, is itself a controversial and poorly understood matter.

A full account of what autonomy consists in, and whether or not machines could possess it, would require answering a set of questions about the nature and limits of the will, which have puzzled philosophers for centuries. This is a task well beyond the scope of the current paper. Instead, I shall have to be content with two observations.

Firstly, a large number of influential writers have argued that future artificial intelligences will possess capacities equal to, and even exceeding those of human beings.[23] Should this occur then these machines will presumably have a strong claim to be autonomous as well.[24] At least some serious writers in the area are therefore committed to the claim that the dilemmas that I study here will arise.[25]

Secondly, that autonomy and moral responsibility go hand in hand. To say of an agent that they are autonomous is to say that their actions originate in them and reflect their ends. Furthermore, in a fully autonomous agent, these ends are ends that they have themselves, in some sense, chosen. Their ends result from the exercise of their capacity to reason on the basis of their own past experience. In both of these things, they are to be contrasted with an agent whose actions are determined, either by their own nature, or by the ends of others. Where an agent acts autonomously, then, it is not possible to hold anyone else responsible for its actions. In so far as the agent's actions were its own and stemmed from its own ends, others can not be held responsible for them.[26] Conversely, if we hold anyone else responsible for the actions

© Society for Applied Philosophy, 2007

of an agent, we must hold that, in relation to those acts at least, they were not autonomous.

For the moment, I want to remain agnostic on the question of the extent to which existing or future AWS can truly be said to be autonomous. Ultimately, I will argue that even 'strong AI's' of the sort discussed by Kurzweil, Brooks, Moravec, and others, are likely to inhabit a 'grey area' in the middle of the range between systems which are determined and those which are full moral agents, precisely because it is difficult to see how it would be possible to hold them responsible for their actions. However, what I want to do here is to take seriously for the moment the possibility that they might exercise a substantial degree of autonomy and see what follows from that. I shall argue that the more these machines are held to be autonomous the less it seems that those who program or design them, or those who order them into action, should be held responsible for their actions. This itself is enough to suggest that the prospects for fully autonomous machines are more remote than is sometimes claimed.

Robot Warriors and Robot War Crimes

The prospect of the deployment of AWS raises many disturbing ethical questions. How will the use of robot weapons affect the ways in which wars are fought, the level and nature of casualties, and the threshold of conflict? What sort of decisions should they be allowed to make? How should they be programmed to make them? Should we grant non-human intelligent agents control of powerful weapons at all? If a system is intelligent enough to be trusted with substantial decisions making responsibility in battle, should it also be granted moral standing? Should they then be granted rights under the Geneva Conventions? These are just some of the difficult questions surrounding the ethics of the use of AWS.[27]

The question I am going to consider here is who should be held responsible if an AWS was involved in a wartime atrocity of the sort that would normally be described as a war crime. The reason I shall concentrate on this question is that it is arguably prior to the others suggested above. If, as I shall argue below, it turns out that no-one can *justly* be held responsible for the actions of these systems, then it will be unethical to use them in war. The other questions then need not arise.[28]

Let us imagine that an airborne AWS, directed by a sophisticated artificial intelligence, deliberately bombs a column of enemy soldiers who have clearly indicated their desire to surrender. These soldiers have laid down their weapons and pose no immediate threat to friendly forces or non-combatants. Let us also stipulate that this bombing was not a mistake; there was no targeting error, no confusion in the machine's orders, etc. It was a decision taken by the AWS with full knowledge of the situation and the likely consequences. Indeed, let us include in the description of the case, that the AWS had reasons for what it did; perhaps it killed them because it calculated that the military costs of watching over them and keeping them prisoner were too high, perhaps to strike fear into the hearts of onlooking combatants, perhaps to test its weapon systems, or because the robot was seeking to revenge the 'deaths' of robot comrades recently destroyed in battle.[29] However whatever the reasons, they were not the sort to morally justify the action. Had a human being committed the act, they would immediately be charged with a war crime.

© Society for Applied Philosophy, 2007

Who should we try for a war crime in such a case? The robot itself? The person(s) who programmed it? The officer who ordered its use? No one at all? As we shall see below, there are profound difficulties with each of these answers.

Responsibility and Jus in Bello

The question of the attribution of responsibility in this situation matters because I take it that it is a fundamental condition of fighting a just war that someone may be held responsible for the deaths of enemies killed in the course of it. In particular, someone must be able to be held responsible for civilian deaths. The responsibility at issue here is moral and legal responsibility and not mere causal responsibility.

This condition may be thought of as one the requirements of *jus in bello*: it may also be thought of as a precondition of applying this idea at all.

It is a minimal expression of respect due to our enemy — if war is going to be governed by morality at all — that someone should accept responsibility, or be capable of being held responsible, for the decision to take their life. If we fail in this, we treat our enemy like vermin, as though they may be exterminated without moral regard at all. The least we owe our enemies is allowing that their lives are of sufficient worth that someone should accept responsibility for their deaths.[30] Similarly, their grieving relatives are entitled to an answer as to why they died, which includes both knowing who is responsible and what their reasons were. Ensuring that someone can be held responsible for each death caused in war is therefore an important requirement of *jus in bello*.

The assumption and/or allocation of responsibility is also vital in order for the principles of *jus in bello* to take hold at all. The principle of discrimination, for instance, which requires that combatants distinguish between legitimate and illegitimate targets, assumes that we can specify who is responsible for attacks that may violate it. More generally, application of the principles of *jus in bello* requires that we can identify the persons responsible for the actions that these principles are intended to govern.

Whichever way we understand it, the principle that we must be able to identify those responsible for deaths in war receives support from strong consequentialist and deontological arguments. As I set it out above, it is a necessary condition of the respect for persons that is at the heart of Kantian, and other deontological, ethics. However, it is also clearly supported by weighty consequentialist considerations. An inability to identify those responsible for war crimes would render their prosecution moot, for instance, with disastrous consequences for the ways in which wars are likely to be fought. Indeed the very same consequentialist reasons which motivate a concern for *jus in bello* in the first place ground our interest in the question of responsibility.

Sometimes, of course, there will be individual circumstances in which it is exceedingly difficult to determine who should be held responsible for certain deaths, or perhaps even where no-one could justly be held responsible. Accidents do happen, and even more so in war. However, these accidents represent regrettable, if inevitable, failures to live up to principles of justice in war fighting. If the nature of a weapon, or other means of war fighting, is such that it is *typically* impossible to identify or hold individuals responsible for the casualties that it causes then it is contrary to this important requirement of *jus in bello*. It will not be ethical to employ this means in war.

Part of what is immoral about weapons of mass destruction, or other means of indiscriminate slaughter, such as anti-personnel mines, for instance, is that they violate

© Society for Applied Philosophy, 2007

this condition. It is traditional to characterise the problem with these weapons as a failure to distinguish between legitimate and illegitimate targets, typically, soldiers and civilians.[31] However, another way of thinking about these weapons is that when they are used, no one is taking responsibility for the decision about who does and does not get killed.[32] It is mere accident that *these* particular civilian deaths occurred. The use of these weapons implies that this is not important and thus demonstrates a profound disrespect for the value of an individual human life.[33]

If it turns out that no one can properly be held responsible in the scenario described above, then AWS will violate this important condition of *jus in bello*.

Will Insisting on 'Human Oversight' Avoid the Problem?

The thought that a machine might be trusted to make the decision to take a human life is obviously a disturbing one. In order to avoid the ethical dilemmas this might pose, or the public outcry it is likely to provoke, we might wish to ensure that any decision to open fire, or to take action that could threaten human life, be considered and approved by a human operator.[34] When I contacted the Program Manager of the US FCS project in the course of researching this paper, he was very quick to insist that this would be the case.[35]

Requiring that human operators approve any decision to use lethal force will avoid the dilemmas described here in the short-to-medium term. However, it seems likely that even this decision will eventually be given over to machines. There is an obvious tension involved in holding that there are good military reasons for developing autonomous weapon systems but then not allowing them to fully exercise their 'autonomy'. The same pressures that are pushing for the deployment of military robots in the first place also push for them to be given control over which targets to attack and when to open fire.

Indeed, as AI technology improves, a human operator may prove not merely redundant but positively disadvantageous in such systems. 'Improvements' in the technology of war fighting mean that the tempo of battle is continually increasing. The development of long range and high speed projectiles, with ever improving guidance and target acquisition systems means that in some combat arenas, especially air combat, the window of opportunity in which to take evasive or effective counter action once hostile contact is made is contracting rapidly. It seems likely that sometime in the not-too-distant future, the time available to make survival critical decisions will often be less than the time required for a human being to make them.[36] When this occurs, then *only* robots will be capable of participating effectively in these forms of combat. Many military commentators have observed that the generation-after-next of fighter aircraft are likely to be uninhabited for precisely this reason. Furthermore, the communications infrastructure required in order to maintain human oversight is an obvious weak point in the operations of these systems.[37] The links between the weapon system and its human overseer may be threatened by electronic counter-measures by hostile forces, by environmental factors (such as difficulties in maintaining line-of-sight necessary for communications, atmospheric interference, etc) or other exigencies of the 'fog of war'. The communications infrastructure itself is likely to be an early target for enemy attack, with the prospect that success in such attacks could disable all robot forces. The survivability and range of operations of an AWS will therefore be greatly increased if it is capable of operating without human supervision.

© Society for Applied Philosophy, 2007

Both the tendency of the tempo of battle to increase with technological developments and the costs associated with keeping a human 'in the loop' are likely to be greatly exacerbated as soon as autonomous weapons systems are deployed. Weapons that require human oversight are likely to be at a substantial disadvantage in combat with systems that can do without. Thus as soon as one nation is capable of deploying AWS that can operate without human oversight then all nations will have a powerful incentive to do so.

For all these reasons, there is likely to be strong pressure in the future to allow AWS to operate in a 'fully autonomous' mode — including the ability to make decisions about choice of targets and payload deployment.

In any case, a version of the problem that concerns me here arises even when human beings have the last word on a decision, as long as human combatants are relying crucially on other decisions made by AI's. For instance, the pilots of modern fighter aircraft must make sense of a tremendous amount of rapidly evolving information from multiple sources, very quickly, in situations where reaching the correct decision is a life or death matter. In order to make this task easier, research is under way into systems in which a computer, using artificial intelligence technology, analyses and interprets the incoming data from various radar and other sensing systems and then re-presents it to the pilot in the form of moving icons indicating the probable nature of various objects (including 'friend or foe' identification), the level of threat they represent and their likely future trajectory. The pilot can then act on the basis of this information.[38] Where such a system is in use, the fighter pilot relies crucially on targeting information provided by an AI in making the decision to destroy a target. If this information turns out to be wrong — perhaps even 'deliberately' misleading — can we still hold the pilot responsible for the consequences of their decision? The question of who is responsible for the decisions on the basis of which the human decision is made, remains crucial.[39]

Responsibility for Robot War Crimes

Who should we hold responsible for a war crime in a situation that crucially involves a decision made by an AWS, such as that described above?

The Programmer?

Given that the weapon is presumably not supposed to behave in this way, it is tempting to insist that the fault lies with the person(s) who designed and/or programmed the weapon, and that they should be held responsible for its destructive result.[40] However, this will only be fair if the situation described occurred as a result of negligence on the part of the design/programming team.

This need not be the case for two reasons.

Firstly, the possibility that the machine may attack the wrong targets may be an acknowledged limitation of the system. If the manufacturers have made this clear to those who purchase or deploy the system, then it seems they can no longer be held responsible, should this occur. The responsibility is assumed by those who decide to send the weapon into battle regardless.

© Society for Applied Philosophy, 2007

Secondly, and more importantly, the possibility that an autonomous system will make choices other than those predicted and encouraged by its programmers is inherent in the claim that it is autonomous. If it has sufficient autonomy that it learns from its experience and surroundings then it may make decisions which reflect these as much, or more than, its initial programming. The more the system is autonomous then the more it has the capacity to make choices other than those predicted or encouraged by its programmers. At some point then, it will no longer be possible hold the programmers/designers responsible for outcomes that they could neither control nor predict. The connection between the programmers/designers and the results of the system, which would ground the attribution of responsibility, is broken by the autonomy of the system. To hold the programmers responsible for the actions of their creation, once it is autonomous, would be analogous to holding parents responsible for the actions of their children once they have left their care.

The Commanding Officer?

The argument above suggests that the officer who ordered the deployment of the weapons system should instead be held responsible for the consequences of its use. The risk that it may go awry is accepted when the decision is made to send it into action. This is the preferred approach of the military forces seeking to deploy existing AWS.[41] It accords with the precedent set by our response to cases where other weapons that may kill people other than their intended target are used. In these cases we simply insist that those who use them should be held responsible for the deaths they cause, even where these were not intended.

If the autonomy of the weapon *merely* consists in the fact that its actions cannot always be reliably predicted and therefore that it may sometimes kill people whose deaths were not intended, then the analogy with existing weapons may be close enough. Employing AWS, then, is like using long-range artillery. The risk that shells may land off target is accepted when the decision to fire is made. If they do kill people other than their intended targets, responsibility for the decision to fire remains with the commanding officer.

However, this is a peculiar way to treat what are advertised as 'smart' weapons. It implies that there is no fundamental moral difference between them and more ordinary 'dumb' weapons. This way of resolving the problem therefore sits uneasily with the original claims about the 'autonomy' of such systems. What distinguishes AWS from existing weapons is that they have the capacity to choose their own targets. If we understand the autonomy they exercise in doing so only as a limit on our ability to predict how they will behave, then on the face of it this implies that the more autonomous they become the less confidence we can have that they will attack the targets that we intend. However, we normally think that smart bombs, and other AWS, are *more* reliable ways to attack the enemy. It is hoped that future AWS will even be capable of discriminating reliably between civilian and military targets. This has even led a number of critics to argue that the use of these weapons is morally *superior* to the use of ordinary 'dumb' weapons.[42]

So, the autonomy of the systems cannot be captured by the mere fact that they are unpredictable. Yet these weapons are more than just guided weapons, which attack targets that have been chosen for them. The more autonomous these weapons are, the

© Society for Applied Philosophy, 2007

more it is possible that they might attack the wrong target. That is, the more it is true that they *could* do so. The precise nature and significance of this 'could' must remain somewhat mysterious. Again, the nature of free will is too large a topic for me to treat here. At this stage of my discussion it will have to suffice to note that the autonomy of the machine implies that its orders do not determine (although they obviously influence) its actions. The use of autonomous weapons therefore involves a risk that military personnel will be held responsible for the actions of machines whose decisions they did not control. The more autonomous the systems are, the larger this risk looms. At some point, then, it will no longer be fair to hold the Commanding Officer responsible for the actions of the machine. If the machines are really choosing their own targets then we cannot hold the Commanding Officer responsible for the deaths that ensue.

The Machine?

The final possible loci of responsibility, then, is the machine itself. Perhaps we should try the machine for war crimes?

It is hard to take seriously the idea that a machine should — or could — be held responsible for the consequences of 'its' actions.[43] We can easily imagine a robot, or for that matter, any machine, being *causally* responsible for some death(s). A number of authors have even recently argued that machines may be properly be thought of *as* acting; that they are 'artificial agents'.[44] However, we typically baulk at the idea that they could be *morally* responsible.[45]

Why should it be so hard to imagine holding a machine responsible for its actions? One reason is that it is hard to imagine *how* we would hold a machine responsible — or, to put it another way, what would follow from holding it to be responsible. To hold that someone is morally responsible is to hold that they are the appropriate locus of blame or praise and consequently for punishment or reward. A crucial condition of the appropriateness of punishment or reward is the conceptual possibility of these treatments. Thus in order to be able to hold a machine morally responsible for its actions it must be possible for us to imagine punishing or rewarding it.[46] Yet how would we go about punishing or rewarding a machine?

For some people, the intuitive implausibility of punishing or rewarding a machine will be enough to establish that the AWS could never be held morally responsible for its actions. However, I want to investigate the matter a little bit further in order to show that what would need to be true in order for us to be able to hold a machine responsible is even more demanding than might at first sight appear. I shall concentrate on the question of punishment as the issues are starker here; but the arguments explored below will also apply to the question of the possibility of rewarding a machine.

Some people may not, in fact, share the intuition that we could not punish a machine, especially when it comes to robots that are sufficiently intelligent to have a good claim to be described as autonomous. It seems likely that any robot that is capable of 'intelligent' behaviour will have internal states which function like desires as well as an internal structure which motivates them to pursue these.[47] These cognitive states will be necessary in order to make it possible that the machine can be rewarded, or experience reward, when it achieves its (or perhaps its designer's) goals. Frustration of these desires might therefore function as a mechanism whereby the machine could be punished.

© Society for Applied Philosophy, 2007

If we imagine an AI with intellectual capacities similar, or even equal to those of human beings, then it seems even more plausible that they could be punished. Such machines will presumably have goals and desires beyond those presupposed by their military role. If an AWS earned wages for its services to the nation, which it could spend on its own projects, then we could simply dock these for minor misdemeanours. For more serious crimes, we could imprison the culprit. One could, for instance, restrict a machine's liberty by restraining it physically or by imposing a restraint through its programming. Alternatively, we might administer corporeal punishment by damaging the machine in some way, or perhaps by administering electric shocks to those electrodes through which it senses damage in combat. Finally, we might institute 'capital' punishment for the most serious crimes, such as war crimes, and destroy the machines responsible for them.

However, although these courses of action might appear to satisfy our psychological need for revenge, it is not yet clear that they would count as punishment. In order for any of these acts to serve as punishment they must evoke the right sort of response in their object. The precise nature of this response will depend on our theory of punishment. I shall assume that the most plausible accounts of the nature and justification of punishment require that those who are punished, or contemplate punishment, should suffer as a result.[48] While we can imagine doing the things described above to a machine, it is hard to imagine it suffering as a result.

In order for a machine to be capable of being punished then, it must be possible for it to be said to suffer. Furthermore, its suffering must be of the sort that we find morally compelling. Talk of machine suffering which is most naturally expressed in inverted commas — 'suffering' — will not suffice here. If a grieving relative of one of the machine's victims questions whether the machine has been punished sufficiently, it will not do to point out that it is 'suffering' because there is friction in its gears as a result of not being oiled, or that it hasn't been able to log on to the web to play chess in its spare time. In order for our treatment of the machine to count as punishment, it must be capable of suffering in ways that might motivate the same set of responses that we have as a matter of course to human beings. It must be such that we could understand someone saying that they felt sympathy for it, or grief, or remorse, if this suffering turned out to be unnecessary.[49] Indeed, the suffering involved in a machine being punished must be such that if we discover that it was in fact innocent (perhaps the bomb that killed the soldiers was launched by another AWS near by) then we feel that we have done it a serious wrong and owe it recompense.

What must be true in order for a machine to be held morally responsible is therefore much more demanding than at first appeared. Not only must machines have internal states of equivalent complexity and with analogous structure to those of human beings, but they must also have the capacity to express these in ways that will establish the moral reality of these states. They must make the same sorts of moral and empathic demands upon us as do other (human) people.

Again it might be thought that robots of the sort predicted by Brooks, Kurzweil, and Moravec would possess sufficient internal complexity and capacity for expressiveness for this to be the case.[50] That is, if we were to see them 'suffering' we would no longer be inclined to feel that we need the inverted commas — we would believe they really *were* suffering. Similarly, we would instinctively reach out to give them assistance, feel concern for their wellbeing, and grieve for their deaths. If, as a result, we felt of such

© Society for Applied Philosophy, 2007

machines that they were appropriate sites of punishment, then they could be held responsible for their actions; but equally well we could be held responsible for our actions in relation to them. They would be full 'moral persons'. Paradoxically, the creation of such machines would not achieve the goal that motivated their development in the first place — that of allowing wars to be fought without risking our soldiers being killed. Our machines would have become our soldiers and we should be as morally concerned when our machines are destroyed — indeed killed — as we are when human soldiers die in war.

I will not attempt to settle here whether or not machines will ever be able to meet these conditions — although it is my belief that they will not.[51] It will serve here to point out the enormous gap between any system that might meet these conditions and any AWS currently in existence, or even on the drawing board. Even a machine that was capable of making exceedingly complex battlefield decisions, in a fraction of a second, on the basis of a knowledge of tactical and technical expertise way beyond that possessed by any human being would not meet these conditions. While some hypothetical AI that can establish a claim to moral personhood might be responsible for what it does, for the foreseeable future we will not be able to hold that machines are responsible for their actions.

We have reached an impasse. I have argued that as machines become more autonomous a point will be reached where those who order their deployment can no longer be properly held responsible for their actions. I have also suggested that machines could have capacities way beyond those necessary to reach this point without it being possible to hold them responsible. How can both these things be true — and how can we proceed from here?

Robot Warriors and Child Soldiers

We can better conceptualize this dilemma if we consider another case where attribution of responsibility in wartime is problematic; the use of child soldiers in war.[52] One of the things that is unethical about employing youth below a certain age in combat duties is that life and death decisions are placed in the hands of children who cannot be held responsible for their actions.[53] While they lack full moral autonomy — and therefore are not morally responsible for what they do — there is clearly *a* sense in which children are autonomous. They are capable of a wide range of decisions and actions. They are certainly much more autonomous than any existing robot. Yet they are not appropriate objects of punishment, as they are not capable of understanding the full moral dimensions of what they do — and therefore of understanding the connection between their punishment and their crime.

The limited autonomy that children do possess is enough, moreover, to ensure that those who order them into action do not control them. The fact that their commanders do not control them problematises the attribution of responsibility for their actions. Note that what makes the attribution of responsibility especially problematic here is not that child soldiers are necessarily unreliable or unpredictable. For instance, it might be a more reliable way to kill a particular enemy to send a child assassin to seek him or her out, than to drop a bomb from great height. Yet while they will *probably* kill the right person, they might not. This possibility is inherent in their capacity for autonomous action.

© Society for Applied Philosophy, 2007

It is the prospect of intelligent actors without any moral responsibility that makes child armies especially terrifying. When child armies take to the battlefield, as they have in Angola and Liberia in recent years, no one is in control. If civilians are killed they are killed senselessly without anyone being responsible for their deaths. The deaths that occur, occur, in a sense, indiscriminately — without necessarily being random.

There seems to be a conceptual space in which children and (perhaps) machines are sufficiently autonomous to make the attribution of responsibility to an appropriate adult problematic, but not so autonomous as to be responsible for their own actions. This space is bounded at the lower end by entities that have no, or at least little, autonomy. If they play a role in generating an outcome, it can only be a causal and not a moral one; if anyone is responsible for that outcome, it is the person who placed them in the position where they played that causal role. This space is bounded at the upper end by entities that are fully morally autonomous, are responsible for their own actions, are appropriate sites of punishment and also therefore make moral claims upon us. In between is a region, fuzzy at both the upper and lower ends, in which entities are sufficiently complex, and possess internal states that function as ends, such that their actions can no longer be attributed to those who set them in motion, but where they are not sufficiently well-formed moral agents to be fully morally responsible.

In practice, I think we try to close this space by stipulating that all entities should be treated as though they fit beneath the lower or above the upper bound. However, as the case of children shows, this is not always a convincing solution.

The problem that I have identified with robot weapons is that they seem likely to occupy this uneasy space for the foreseeable future. I have argued that it is easy to imagine autonomous machines entering this space from their current place well below its bottom limit. But is difficult to imagine them emerging from the upper limit of this space.

While they remain in this ambiguous space, the attribution of responsibility for their actions is deeply problematic. The only possible solution seems to be to *assign* responsibility to an appropriate individual — presumably the commanding officer who orders their use. However, as I have argued above, this solution holds the commanding officer responsible for things that they could not control, and therefore risks that they will be punished unfairly. The only way of meeting our obligation to enemy combatants to ensure that someone can be held responsible if they are killed unjustly, risks a grave injustice to our own military personnel who wield authority on the battlefield.

Conclusion

I have argued that it will be unethical to deploy autonomous systems involving sophisticated artificial intelligences in warfare unless someone can be held responsible for the decisions they make where these might threaten human life. While existing autonomous weapons systems remain analogous to other long-range weapons that may go awry, the more autonomous these systems become, the less it will be possible to properly hold those who designed them or ordered their use responsible for their actions. Yet the impossibility of punishing the machine means that we cannot hold the machine responsible. We can insist that the officer who orders their use be held responsible for their actions, but only at the cost of allowing that they should sometimes be held entirely responsible for actions over which they had no control. For the

© Society for Applied Philosophy, 2007

foreseeable future then, the deployment of weapon systems controlled by artificial intelligences in warfare is therefore unfair either to potential casualties in the theatre of war, or to the officer who will be held responsible for their use.

Robert Sparrow, School of Philosophy and Bioethics, Faculty of Arts, Monash University, Victoria 3800, Australia. Robert.Sparrow@arts.monash.edu.au

NOTES

1 Indeed, the military is one of the major sources of funding for research into robotics and artificial intelligence. In a recent survey of cutting edge research in robotics, there is scarcely a laboratory in Europe or the US that is not receiving funding from the military. See P. Menzel, and F. D'Aluisio, *Robo Sapiens: Evolution of a New Species* (Cambridge, MA: MIT Press, 2000).
2 C. Beal, 'Briefing Autonomous Weapons Systems: Brave New World', *Jane's Defence Weekly*, 33, 6 (2000): 22–26.
3 'Low Cost Autonomous Attack System (LOCAAS)', http://www.fas.org.man/dod-101/sys/smart/locaas.htm, at 14.8.02.
4 T. Barela, 'Anti-Armor Einstein' *Airman Magazine* September (1996); Air Armament Center Public Affairs Report 'This bomb can think before it acts' *Leading Edge Magazine*, 42, 2 (2000): 12.
5 P. Meilinger, 'Precision Aerospace Power, Discrimination, and the Future of War', *Aerospace Power Journal* 15, 3 (2001): 12–20.
6 Information on US Navy UUV programs is available via, 'Autonomous Operations', http://www.onr.navy.mil/auto-ops/introduction.asp, at 5.8.02. See also 'Ocean systems — Surveillance & Reconnaissance Systems', http://www.boeing.com/defense-space/infoelect/lmrs/sld001.htm, at 5.8.02.
7 Office of the Secretary of Defense, United States Government (2001) *Unmanned Aerial Vehicles Roadmap 2000–2025* (Department of Defense, United States Government, Washington D.C. Available at www.acq.osd.mil/usd/uav_roadmap.pdf), p. i.
8 R. Chapman II, 'Unmanned Combat Aerial Vehicles: Dawn of a New Age?' *Aerospace Power Journal* 16, 2 (2002): 60–73.
9 S. Shactman, 'Attack of the Drones' *Wired Magazine* 13, 6 (2005) available at http://www.wired.com/wired/archive//13.06/drones_pr.htmal, at 25.08.05.
10 S. Kainikara, 'UCAVs probable lynchpins of future air warfare', *Asia-Pacific Defence Reporter*, 28, 6 (2002): 42–45.
11 Kainikara op. cit. See also, J. Kerr, 'Boeing looks to the future' *Asia-Pacific Defence Reporter*, May 28, 4 (2002): 24–25. The cost savings offered by UCAVs are subject to some controversy but are expected to flow from the relative expense of pilot training, support and training mission related maintenance over the operational life of manned systems. Office of the Secretary of Defense, United States Government (2001) *Unmanned Aerial Vehicles Roadmap 2000–2025*.
12 M. Behar, 'The New Mobile Infantry' *Wired* May, 10, 5 (2002): 110–114; B. Bender 'USMC to buy robots for urban warfare operations', *Jane's Defence Weekly*, 7 June 33, 23 (2000): 2; D. Lewin, 'Smart cars in peace and war' *IEEE Intelligent Systems* May–June, (2001): 4–8.
13 'US plans "robot troops" for Iraq', BBC News World Edition Website, 23 January (2005), available at http://news.bbc.co.uk/2/hi/americas/4199935.stm, at 25.9.05.
14 'Marine Corps Warfighting Laboratory', http://www.mcwl.quantico.usmc.mil/tech.html, at 2.8.02.
15 Descriptions of the FCS program may be found at; 'TTO Programs — Future Combat System (FCS)', http://www.darpa.mil/tto/programs/fcs.html, at 5.8.02; http://www.darpa.mil/darpatech2000/speeches/TTOspeeches/TTOFutureCombat(Van%20Fosson).doc, at 5.8.02; 'Future Combat Sytems/Future Combat System', http://www.fas.org/man/dod-101/sys/land/fcs.htm, at 5.8.02. See also, J. Kerr, 'Boeing looks to the future' (2002).
16 D. Wastell, 'Robot soldiers herald the era of cheaper, "safer" wars', *The Age*, Melbourne, Australia, March 4, (2002).
17 In particular, existing UAVs and UCAVs rely, for at least some of their operations, on a human operator piloting them remotely via satellite data link.
18 Beal op. cit. Note that LOCAAS is being designed to make these decisions, for instance.

© Society for Applied Philosophy, 2007

The responsibility gap: Ascribing responsibility for the actions of learning automata

Andreas Matthias, Ethics and Information Technology 6: 175–183 · 2004
https://doi.org/10.1007/s10676-004-3422-1

Ethics and Information Technology **6:** 175–183, 2004.
© 2004 *Kluwer Academic Publishers. Printed in the Netherlands.*

The responsibility gap: Ascribing responsibility for the actions of learning automata

Andreas Matthias
Computing Centre, University of Kassel, D-34109 Kassel, Germany
E-mail: matthias@hrz.uni-kassel.de

Abstract. Traditionally, the manufacturer/operator of a machine is held (morally and legally) responsible for the consequences of its operation. Autonomous, learning machines, based on neural networks, genetic algorithms and agent architectures, create a new situation, where the manufacturer/operator of the machine is *in principle* not capable of predicting the future machine behaviour any more, and thus cannot be held morally responsible or liable for it. The society must decide between not using this kind of machine any more (which is not a realistic option), or facing a responsibility gap, which cannot be bridged by traditional concepts of responsibility ascription.

Key words: artificial intelligence, autonomous robots, learning machines, liability, moral responsibility

Introduction

When people act, their actions have an impact on the lives of others; and so human societies, in the course of the centuries, have developed elaborate and differentiating rule systems in order to ascribe the responsibility for an action and its consequences justly.[1] When we judge a person to be responsible for an action, we mean either that the person should be able to offer an explanation of her intentions and beliefs when asked to do so, or that, following Strawson (1962), the person is *rightly* subject to a range of specific reactive attitudes like resentment, gratitude, censure, or praise (Oshana 2002: 263). For a person to be *rightly* held responsible, that is, in accordance with our sense of justice, she must have *control* over her behaviour and the resulting consequences "in a suitable sense" (Fischer and Ravizza 1998: 13). That means that the agent can be considered responsible only if he knows the particular facts surrounding his action, and if he is able to freely form a decision to act, and to select one of a suitable set of available alternative actions based on these facts.

Regarding the consequences of the operation of machines, we usually ascribe the responsibility for them to the operator of the machine, as long as the machine operates as specified by the manufacturer. The operator, by putting the machine into operation according to the manufacturer's specification, signals her acceptance of this responsibility. In case the machine does not operate according to the manufacturer's specification (that is, in case it has a flaw in its construction), we ascribe the responsibility to the manufacturer of the machine instead of the operator. This is in accordance with the principle mentioned above of *control* being a necessary condition of responsibility.[2] The operating manual of a device transfers control of that device from the manufacturer to the operator, by specifying the precise set of actions and reactions (in a system theory vocabulary: of *transformations*) the device is expected to undergo during normal operation,[3] thus enabling the operator

[1] For the purpose of this discussion, we will not usually need to distinguish between moral and legal responsibility (including liability). They will be dealt with together, and an explicit distinction will only be made where necessary.

[2] Although the notion of control used here is inspired by Fischer and Ravizza, we will not touch on the subject of determinism. This paper is about the practical problems which a society faces when it is trying to ascribe responsibility, and these are not altered by the determinism discussion, whatever its outcome may be.

[3] The *'operating manual'* may be quite an implicit affair when common artifacts of everyday life are concerned, for example, candles. There is (in the framework of a given society) a nearly universal understanding about how a candle is to be operated and what transformations it may undergo during operation. If, while using it properly, the operator burns down his house, then he himself is held fully responsible. If, on the other hand, the candle *explodes* while burning, then this behaviour is considered not to be part of the ordinary set of candle-usage transformations, and so responsibility is ascribed to the manufacturer instead, because the (implicit) set of operation instructions for a candle does not include provision for the case of an explosion, and thus the operator is bound to have *reduced control* over the device, which results in reduced (or absent) responsibility.

to handle it in a predictable manner, according to her own decisions on how to act.

In situations where the operator has reduced control over the machine he also bears less or no responsibility. If, for example a NASA technician operates from earth a remotely controlled Mars vehicle, and due to bad visibility in a Mars sandstorm and long response times (radio signals take 20 minutes to travel from Earth to Mars) the vehicle falls into a hole and is lost, then we would not consider the technician responsible for the loss. But who can be held responsible instead? In fact, *nobody*. In such cases of accidents that occur through no fault of a specific person, society refrains from ascribing responsibility, and collectively bears the cost resulting from the accident's consequences.[4]

In the following sections of this paper we will see how certain recent developments in the way of manufacturing computerised, highly adaptive, autonomously operating devices, inevitably lead to a partial loss of the operator's control over the device. At the same time, the degree in which our society depends on the use of such devices is increasing fast, and it seems unlikely that we will be able or willing to abstain from their use in the future. Thus, we face an ever-widening *responsibility gap*, which, if not addressed properly, poses a threat to both the consistency of the moral framework of society and the foundation of the liability concept in law.

Applications of learning automata

In order to provide the argument with some technical substance, let us look at a few examples: systems that are in development or already in regular use today.[5]

1. Let us first revisit the Mars vehicle case. Like the NASA *Pathfinder* (Morrison and Nguyen 1996), the vehicle will not only be controlled remotely from Earth, but it will also have its own integrated navigation and control system that enables it to avoid obstacles autonomously. Unlike *Pathfinder,* we will assume that the control program learns: after crossing a certain stretch of terrain, it will store into its internal memory an optical representation of the terrain as a video image, together with an estimate of how easy it was to cross that particular type of terrain. When a similar video image appears next time, the machine will be able to estimate the expected difficulty of crossing it, and it will thus be able to navigate around it if this seems desirable.
   Now let us assume that the vehicle again falls into a hole. Who is responsible? The operator obviously is not, since what caused the vehicle to disappear into the hole was not a manual control command, but a decision taken by the machine's control software itself. So, can the programmer be held responsible? He can deny responsibility by saying that the algorithm used was appropriate and correctly implemented. Surely one can assume that similar kinds of terrain will give similar video image representations. The actual decisions of the control program were based not only on preprogrammed data, but on facts that were added to the machine's database only *after* it reached the surface of Mars: they are not part of the initial program, but constitute genuine experience acquired autonomously by the machine in the course of its operation.
2. In high-rise office buildings, adaptive elevator systems already are used regularly.[6] These systems analyse traffic patterns, typically using artificial neural networks and reinforcement learning algorithms, and they try to minimise waiting and transportation time for the users of the elevator. Now let us suppose, that such a learning, adaptive elevator leaves an important executive waiting for half an hour in the 34th floor, so that he cannot attend a business meeting at which he is expected. A considerable financial damage is caused. Who can be held responsible?[7] The manufacturer can deny responsibility because the elevator, being able to learn, had changed the parameters of its program during the course of its operation, so as to better adapt to the traffic patterns in the building. Because of this, it was no longer possible for the manufacturer to predict or control the specific behaviour of the elevator in a given situation (though it might be possible to prove mathematically that eventually the used algorithm will

[4] Taxation and insurances are some of the ways society has devised to distribute the cost of such accidents to a broad base of its members.

[5] There are many other such systems in use, far too numerous to be dealt with here explicitly, including: electronic noses for banana ripeness determination (Llobet et al. 1999), collision avoidance systems in automatic submarine navigation (Schultz 1991), autonomously flying (Stancliff and Nechyba 2000), game playing (De Jong and Schultz 1988), and web document filtering (Zhang et al. 2000) machines and programs.

[6] OTIS (2003), Schindler (2003), Sasaki et al. (1996).

[7] The case is legally more or less interesting, depending on product liability laws in different countries and legal traditions. But even where no legal liability can be identified (for example in German law), it is possible to ask who *should* (on moral grounds) be considered responsible for the damage done by the machine.

converge to some optimal behaviour, that is, minimal overall waiting time for the users of the elevator). And since the manufacturer could not have predicted or averted the undesirable outcome in this case, he cannot justly be held responsible for this specific behaviour of the machine.[8]

3. Zhou et al. present a system for the automatic diagnosis of lung cancer. The system learns to identify cancer cells on microscope images of specimens of needle biopsies obtained from the bodies of the persons to be diagnosed. Systems like that are supposed to be used where an experienced senior pathologist is not available (e.g., in underdeveloped rural areas). The system has been constructed so that false negative diagnoses are highly improbable (proclaiming the patient to be healthy when there are, in reality, cancer cells present), but there is accordingly much less precaution about false positiva. Although wrong positive diagnoses are not immediately life-threatening, they can cause great financial, practical and emotional problems to the affected parties, and so again we must find someone to ascribe responsibility to for such erroneous diagnoses of the machine.[9] As far as the programmer is concerned, he has done everything possible to prevent false negative diagnoses, which is the best he could do,[10] so he can also not justly be held responsible if the limitations of the machine are clearly known beforehand.[11]
4. One last example: since 1999, there has been a mobile robot distributed by Sony ('AIBO'), which is supposed to be an intelligent toy for children as well as a pet replacement for the small urban apartment. This machine is capable of learning: in Hornby et al. (1999) (and many other publications) we see how AIBO learns new things: words to which it is to react ('come here'), but also new kinds of movements, for example when the built-in stepping algorithm proves not to be optimal for the crossing of deep persian carpets.[12] Let us now assume that an advanced version of the robot is able to change its walking style with time to optimise overall performance. With a little experimentation it will be able to find out that its battery life can be prolonged by galloping, which reduces the friction between itself and the ground. Let us also assume, that the robot, while running around the apartment, collides with a small child and injures him. Who is now responsible? The manufacturer? Why exactly? The child's parents for putting the pet robot into operation in their apartment? Or is this an *unforeseeable* development, which occurred due to the adaptive capabilities of the robot, so that nobody can be justly said to be responsible?

Let us summarise the points: presently there are machines in development or already in use which are able to decide on a course of action and to act without human intervention. The rules by which they act are not fixed during the production process, but can be changed during the operation of the machine, *by the machine itself*. This is what we call machine learning. Traditionally we hold either the operator/manufacturer of the machine responsible for the consequences of its operation, or 'nobody' (in cases, where no personal fault can be identified). Now it can be shown that there is an increasing class of machine actions, where the traditional ways of responsibility ascription are not compatible with our sense of justice and the moral framework of society because nobody has enough *control* over the machine's actions to be able to assume the responsibility for them. These cases constitute what we will call the *responsibility gap*.

In order to fully understand the nature and implications of this problem, it will be necessary to have a look at how learning automata are being constructed.

[8] He might nevertheless be held responsible for selling elevators of this type, if the dangers of using them are not advertised properly in the operating manual. But this is a different point, and, as we will see later on, there might not have been any choice of an alternate, 'safe' system at all.

[9] The trivial solution would be to insist that a human expert should verify the correctness of the machine-made diagnoses and assume the responsibility for them. But this is obviously not possible, because the machine was developed exactly for those situations where a human expert is not available. Were it possible to supply every machine with a controlling human expert, nobody would need the machine in the first place.

[10] When in doubt, the system will always assume that a cancer cell is present. There is no unfailing machine (as there is no such human) and the programmer has (correctly) decided to shift the probability of an error into the less dangerous realm of false positive diagnoses instead of false negative ones.

[11] Here the society as a whole decides to employ an inherently risky technology, because it is assumed that the overall benefits of using it outweigh the risks.

[12] Some of these experiments require modified versions of the robot that are not available on the market, and so does our thought experiment. But this is not relevant to our discussion, since the required modifications can be done easily, at least by the manufacturer himself (and have already been done for experimental purposes).

Engineering learning automata

This is, of course, not the place to give an introduction to artificial intelligence.[13] We will only look at those specific properties of artificial learning systems which are of importance to the question that concerns us, and so, the following exposition will be kept very brief.

We can distinguish four primary types of learning automata which are of interest in the present context: symbolic systems, connectionist architectures (including reinforcement learning systems), genetic algorithms (including genetic programming), and autonomous agents (mobile and immobile).

Symbolic systems

Symbolic systems are built on the assumption that some of the cognitive functionality of humans which we call intelligence can be expressed as a syntactic manipulation of linguistic symbols.[14] A system built along these lines contains long sequences of axioms and derivation rules that are usually expressed in some kind of predicate calculus. The system is either able to derive conclusions from the facts stored in its database, or else to extend this database by adding new rules and facts to it (which is what constitutes learning).[15]

The symbolic approach to artificial intelligence has a long tradition, and it has led to practically usable programs: expert systems that are able to advise users and provide solutions in specific, clearly defined application domains. Classical examples of expert systems include medical diagnosis programs and fault identification in complex technical systems (see Figure 1).

IF:	The exhaust is smoky, and The car is backfiring, and There is a lack of power
THEN:	The carburetor fuel mix is too rich.
IF:	There is a lack of power, and There is a gray deposit on the spark plugs, and The engine overheats,
THEN:	The carburetor fuel mix is too weak.
IF:	The carburetor fuel mix is too rich, or The carburetor fuel mix is too weak,
THEN:	The carburetor fuel mix needs to be adjusted.

Figure 1. Example of an expert system rule base: some rules used in the diagnosis of motor faults (after Andriole and Hopple 1992).

For our present discussion it is important to note that the knowledge of expert systems (and therefore also the actions of such systems, which are based on that knowledge) are stored inside the system in the form of explicit, distinct, quasi-linguistic symbols. They can be inspected at any time and, should need arise, be corrected.

Connectionism and neural nets

While symbolic artificial intelligence presupposes the existence of clear and distinct symbolic representations of objects and the relations between them, connectionism does not. Instead, it attempts to emulate the basic principles of neural operation in living systems. It is based on the observation that biological information processing systems do not seem to represent symbols as discrete entities, but distributed all over the neural net. Information is stored by modifying the architecture of the network and the strength of individual connections between neurons (represented as input 'weights' in artificial networks; see Figure 2). During learning, the network adapts these weights in order to minimise the difference between the actual and the expected output for a given learning pattern. It is essential to this concept that there nowhere is a 'list' or 'catalog' of all learned information, as there is in symbolic programs. Symbolic programs contain axioms, predicates, facts and rules, which represent information *explicitly* and *verifiably*. Connectionist systems lack an explicit representation, and the contained information can only be deduced from their behaviour. Neural networks are, owing to the principles of their architecture, black boxes. We can evaluate their behaviour by applying test patterns and observing their output, but we cannot

– have a look at the information that is stored inside the network, and, even more importantly;

[13] Although we will occasionally use the term 'artificial intelligence' to refer to an academic field of study, we will not enter the discussion as to whether it constitutes "real" intelligence or not. The discussion about the nature of intelligence and if artificial systems can possibly be intelligent is not relevant to the question at hand, since the responsibility gap concerns the *human* part of the human–machine relationship.

[14] These need not necessarily be symbols of a natural language. Often symbolic AI systems incorporate custom formal languages, in which the system's knowledge of the world can be expressed in some way that is useful to the system.

[15] These processes can get quite complex, including programs that derive conclusions by deduction, induction, reasoning by similarity and analogy, conflict resolution between facts in the database, and probabilistic or fuzzy reasoning in the face of unsure facts.

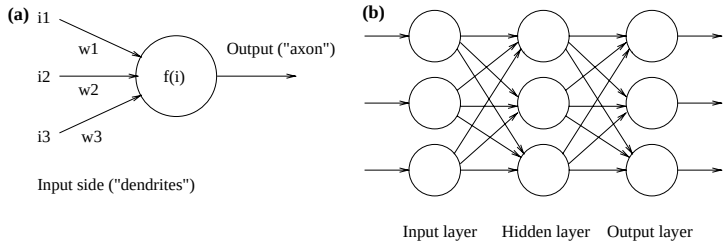


Figure 2. Neural networks.(a) The basic structure of a single artificial neuron. The input signals (i1–i3) are multiplied with (changing) weight factors (w1–w3) and then used as parameters to a function that calculates the output of the neuron. During the learning process these factors change, thus modifying the neuron's behaviour. (b) In a basic backpropagation neural network there are multiple layers of neurons which are interconnected, so that the output of one layer serves as input to the next.

– see what information is *not* represented inside it.[16]

The behaviour of neural networks is not programmed in the way a procedural program is. Instead they are trained *by example:* sets of input patterns are presented to the network and the internal weights are changed gradually, so that the network eventually produces a desired, corresponding set of output patterns. It is not necessary for the trainer of a neural network to be able to express the information to be learned in the form of clear and distinct symbols, so that finally neural networks can successfully learn to make distinctions for which the human trainer himself is unable to provide an algorithmic representation.[17]

Reinforcement learning (Moriarty et al. 1999) lifts the distinction between the training and application phases which we find in traditional neural network concepts. The system learns inside its final operating environment by exploring available action alternatives in a trial-and-error fashion and optimising its own parameters according to the results. Thus, the exploration phase is an integral part of the design of the working machine and cannot be separated from it. This is necessary in highly dynamic environments, where a system needs to change continually in order to achieve optimal efficiency. Consider, for example, a high-rise office tower with ninety floors and a sixteen elevator system which is controlled by software. This system cannot work optimally if it is not always learning and adapting to the changing traffic flows in the building. Imagine there is, for the duration of one week, a conference on floor 34. The system must be able to optimise its behaviour during this time, for example by leaving idle elevators waiting at this floor. After the conference is over, however, this behaviour will have to change again to accommodate new needs. This can only be achieved with systems which learn in their respective operating environments. And since learning is done by trial and error, we find that *errors are unavoidable features of any reinforcement learning system*. This is quite contrary to the common, traditional understanding that technology, done correctly, must operate free of errors, and that errors are always the errors of the programmer, not of the programmed machine. Reinforcement learning presents us, for the first time, with the *necessity* of errors as we know it from living systems: as a system *feature,* the precondition for learning and adaptive behaviour, and not merely a product flaw.

Genetic algorithms

While neural networks are modelled after the neural infrastructure of living systems, *genetic algorithms* (Holland 1975) imitate the principle of evolution through variation, genetic recombination, and selection. In a genetic algorithm system the solution to a problem is typically represented by an ordered chain of symbols that are selected out of an 'alphabet'. Think, for example, of a program that is supposed to find its way through a maze. A solution to the problem of traversing the maze would consist of a sequence of direction change commands: 'right,' 'left,' 'left,' 'right,' and so on, which together describe the way from the entrance to the exit of the maze (see

[16] All the social procedures that humans and animals have developed to evaluate other individuals (rituals of introduction, teaching by demonstration, examinations and psychological tests) are attempts to solve this problem: that we cannot look into the heads of others and *see* the knowledge, intentions and beliefs stored therein. Instead we are forced to infer them indirectly through observation of the other's behaviour and the application of test patterns.

[17] This is especially true when we consider noisy or distorted input, as in the case of optical character recognition, which is very hard to do algorithmically, but almost trivial with artifical neural networks.

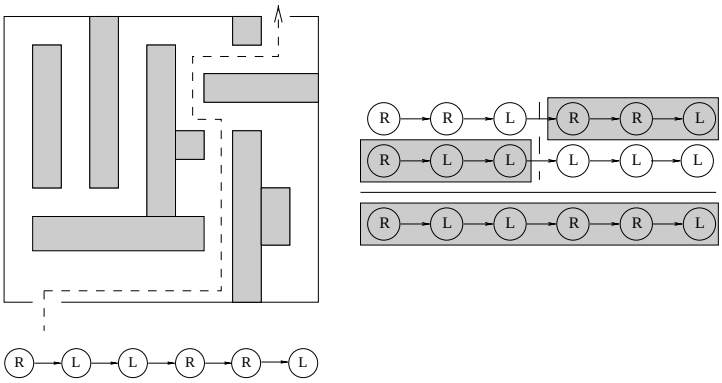


Figure 3. The way through a maze and its representation in a quasi-genetic 'alphabet' of direction-change statements (left). The chain of left–right-statements (the 'genetic material') is manipulated with operations like the cross-over shown here, so that eventually the algorithm arrives at a solution (right).

Figure 3). In the world of the genetic algorithm metaphor, we describe this sequence as the 'genome', which is composed of directional genes with the alleles 'right' and 'left'. The search for a solution begins with a big population of virtual organisms which are initialised with a random genome, that is, with a random sequence of direction-change statements. Limiting factors in the simulated environment provide the required selection pressure: if an organism has been moving around unsuccessfully for a while, or has collided too often with the walls of the maze, it is removed from the simulation (it 'dies') and its genetic information is lost. Successful organisms (that move forward without colliding with the walls or moving in circles) are given the chance to 'reproduce'. During reproduction the genetic material of one or more parental organisms is inherited by the filial organisms.[18] After a while, the simulation produces organisms which can cross the maze successfully (though the solutions found are not necessarily optimal).

Genetic algorithms can be best applied to problems which solutions are representable as (not overly long) chains of symbols, and for which the solution can be approximated gradually by trial and error. Some applications are the evolution of behaviours for mobile robots (Schultz 1994), the classification of sensoric input (Stolzmann et al. 2000), and the navigation of autonomous vehicles (Schultz 1991).

What we have seen just now could be addressed as the "symbolic" variant of genetic algorithms. But it is also possible to combine genetic algorithms with connectionist architectures, by *evolving* neural nets, instead of defining their architecture by hand (Belew, et al 1990; Nolfi and Parisi 1991).

Genetic programming still goes one step further. Here, the result of the simulated evolution is not the solution vector itself any more, but a *software program* that implements the solution. Thus, the genetic algorithm acts itself as a programmer. We will address the significance of this fact in the context of responsibility ascription later in this paper.

Autonomous agents

Finishing this brief survey of artificial intelligence programming methods, we will have a parting look at the technology of artificial agents. Autonomous agents are artificial entities that fulfil a certain, often quite narrow purpose, by moving autonomously through some 'space' and acting in it *without human supervision*. The agent can be a software program that moves through information space (e.g., an internet search engine spider), but it can also have a physical presence (e.g., a computer-driven vacuum cleaner or a robotic pet) and move through the space of an apartment.

Two points concerning agents are interesting in the present context: First, that agents are *per definitionem* designed to act, and that, in the course of their operation, they must inevitably interact with other things, people, and social entities (laws, institutions and expectations). Second, that agents which have a physical presence constitute a new category of machines: such that can learn from the direct interaction with a real environment and that can in return directly manipulate this same environment.

[18] The specific rules of reproduction and genetic material transfer vary widely between different implementations of the genetic algorithm, but this makes no difference for the purpose of this paper.

These machines have (contrary, for example, to desktop computers) an unmediated access to sensory impressions, their symbolic representation, and subsequent actions that lead to new sensory impressions, and they act in the same environment as humans do.

The responsibility gap

Let us now return to the central question of this paper.

To what extent is it possible to hold the manufacturer/programmer/operator of an autonomous, learning automaton responsible for the actions of the machine?

If we look at the role of the machine's manufacturer (or, in our case, the programmer) and the change of this role, from the straightforward programming concepts of traditional software engineering (see Figure 4), to the construction of autonomous artificial intelligence systems, we can clearly see the profound change that has taken place in the past few years. In the beginning we find the programmer as *coder,* that is, someone who expresses the program (and thus the operating behaviour of the machine) line by line and statement by statement in a linguistic representation that can be executed directly by the machine (the statements of a programming language). At any moment, it is possible to inspect the memory of the machine and to determine the precise extent of the program. Changes can be made directly through the addition, removal or exchange of particular statements in the program. The programmer, *in control* of the behaviour of the machine in every single detail, can explain the way his algorithm works, and an observer can follow this explanation and check its correctness. Errors are always errors of the *programmer,* not errors of the *program*. They can always be identified, isolated and fixed, and the programmer can rightly be held responsible for any misbehaviour of the machine.

As the techniques of artificial intelligence programming develop further, the role of the programmer changes. This change is not sudden but gradual, and its extent differs according to the technology employed.

It begins with the programmer who uses logic-oriented programming languages and symbolic expert systems losing control over the execution flow of the program. Programming languages based on predicate logic are not executed in the same linear fashion as their procedural, imperative counterparts: instead of being executed from the first to the last 'command,' they use a run-time system which searches (in a way not always known to the programmer) for deductions from given axioms and inference rules. The flow of control in such systems is typically difficult to describe and should, ideally, be of no interest to the programmer. This is, in itself, not a big problem because as long as there is a symbolic representation of facts and rules involved, we can always check the stored information and, should this be necessary, correct it.

In artificial neural networks, the symbolic representation of information and flow control disappears completely: instead of clear and distinct symbols we have a (possibly very large) matrix of synaptic weights, which cannot be interpreted directly any more. The knowledge and behaviour stored in a neural network can be only inferred indirectly through experimentation and the application of test patterns after the training of the network is finished (as is the case with living systems).

```
FilenameStruct * readdirectory( FILE *fp, int *count) {
          FilenameStruct *fn;
          int i, len = GetWord(fp);

          *count = (len / 17) - 1;
          fn = (FilenameStruct *) malloc(*count * sizeof(FilenameStruct));
          for (i = 0; i < *count; i++) {
               fn[i].offset = GetLong(fp);
               fread(fn[i].fname, 13, 1, fp);
               fn[i].fname[13] = 0;
               lowerstr(fn[i].fname);
               if (showdirectory)
                    fprintf(stderr, "%13s %6d\n", fn[i].fname, fn[i].offset);
               }
          return (fn);
     }
```

Figure 4. Part of a program in the 'C' programming language. The programmer of such a *procedural* or *imperative* program must enter every single program statement herself. All program statements together implement an algorithm for the solution of the problem at hand.

Reinforcement learning, being usually based on neural network concepts, shares the same problems, but additionally it lifts the distinction between a training and a production phase. Reinforcement learning systems explore their action space, while working in their operating environment, which is their central feature (enabling them to adapt to ever-changing environments) as well as a big drawback concerning their predictability. The information stored in the network cannot be fully checked, even indirectly, because it always changes. Even if we can prove mathematically that the overall performance of such a system will eventually converge to some optimum, still there will be *unavoidable* errors on the way to that optimised state. The creator of such a system (not really being a programmer in the traditional sense any more) cannot go about eliminating errors; instead he must explicitly permit them, so that the system can stay operational and improve its performance.

An additional layer of obscurity comes with the use of genetic programming methods. Here, the result of the simulated evolution is not the solution itself, but a program that implements the solution. Thus, we get an additional layer of machine-generated code that gets between the programmer and his product. While in neural networks the designer still defines the operating parameters of the system (the network architecture, the input and output layers and their interpretation); and while with genetic algorithms at least he can define the alphabet used and the semantics of its symbols; with genetic programming he loses even this minimal amount of control, and he creates a machine that programs itself.

Finally, autonomous agents also deprive him of the spatial link between him and his product. The agent acts outside the observation horizon of its creator, who, in the case of a fault, might be unable to intervene manually (because he might not know about the fault until a much later point in time). This is the case for pure information agents (internet indexing programs) as well as for agents which have a physical manifestation (autonomous space vehicles, targeting systems of military missiles, mobile electronic pets).

Thus, we can identify a process in which the designer of a machine increasingly loses control over it, and gradually transfers this control to the machine itself. In a steady progression the programmer role changes from *coder* to *creator* of software organisms.

In the same degree as the influence of the creator over the machine decreases, the influence of the operating environment increases. Essentially, the programmer transfers part of his control over the product to the environment. This is particularly true for machines which continue to learn and adapt in their final operating environment. Since in this situation they have to interact with a potentially great number of people (users) and situations, it will typically not be possible to predict or control the influence of the operating environment.[19]

Conclusion

We presented some of the reasons why certain classes of autonomously adaptive machines must inevitably display suboptimal or even outright erroneous behaviour; behaviour that must be attributed to the machine itself and not to its designer or operator.

1. In the course of the progression of programming techniques: from the conventional procedural program, via neural network simulations, to genetically evolved software, the programmer loses more and more of her control over the finished product. She increasingly becomes a 'creator' of 'software organisms', the exact coding of which she does not know and is unable to check for errors.
2. The behaviour of the machine is no longer defined solely by some initial, and henceforth fixed program, but increasingly shaped by its interaction with the operating environment, from which the machine adapts new behavioural patterns that constitute solutions in the machine's problem space.
3. In order to be able to adapt flexibly to new situations (which is necessary if they operate in dynamic and changing environments), automata must leave behind the clear separation between programming, training, and operation phases. Practically useful technologies will have to learn during operation, which also means that they will have to make 'mistakes' during operation (a ''mistake'' being just the exploration of the solution space by the machine itself, which enables it to arrive autonomously at new solutions).
4. There are an increasing number of situations in which the supervision of an operating machine by a human expert is either in principle or for economic reasons impossible. The supervision of an operating machine is impossible when the machine has an informational advantage over the operator (e.g., navigation computers in cars, radar-based

[19] Except statistically, but a statistical fact (97% of elevators operate within 1% of their optimal performance) is not a guarantee against errors of the machine, and therefore not a useful answer to the responsibility question.

flight control systems, a lung-cancer diagnosis machine in a remote rural area). It is also impossible when the machine cannot be controlled by a human in real-time due to its processing speed and the multitude of operational variables involved (e.g., the control system for a group of 16 elevators in a 90-floor building, or the controlling computer for a nuclear power plant).

5. Still, we cannot do without such systems, because the pattern processing and systems control tasks that we must accomplish in our highly dynamic and complex environments are so complicated that they cannot be addressed by simpler, statically programmed machines. Planetary exploration, traffic control, automated medical diagnosis: these are tasks that cannot be solved by simple, verifiable algorithms. Not to address such tasks seems not to be an acceptable option to most citizens today.

Automatic machines leave their traditional operating environments, and increasingly move into problem areas that, owing to their dynamic nature and complexity, have previously only been accessible to humans. It is natural that in the course of this transition they will not be able to avoid also acquiring some of man's limitations (e.g., learning from experience as the basis of flexible behaviour; experience being just another word for potential error.)

If we want to avoid the injustice of holding men responsible for actions of machines over which they *could not have* sufficient control, we must find a way to address the responsibility gap in moral practice and legislation. The increasing use of autonomously learning and acting machines in all areas of modern life will not permit us to ignore this gap any longer.

References

S.J. Andriole and G.W. Hopple. *Applied Artificial Intelligence: A Sourcebook*. McGraw-Hill, New York, 1992.

R.K. Belew, J. McInerney and N.N. Schraudolph. Evolving Networks: Using the Genetic Algorithm with Connectionist Learning. Cognitive Computer Science Research Group, Computer Science & Engineering Department (C-014), University of California at San Diego. *CSE Technical Report #CS90–174,* June, 1990.

K.A. De Jong and A.C. Schultz. Using Experience-Based Learning in Game Playing. *Proceedings of the Fifth International Machine Learning Conference,* (pp. 284–290). Ann Arbor, Michigan, June 12–14, 1988.

J.M. Fischer and M.S.J. Ravizza *Responsibility and Control. A Theory of Moral Responsibility*. Cambridge University Press, Cambridge, 1998.

J. Holland. *Adaptation in Natural and Artificial Systems*. University of Michigan Press, Ann Arbor, 1975.

G.S. Hornby, M. Fujita, S. Takamura and others. *Autonomous Evolution of Gaits with the Sony Quadruped Robot*. Group 1, D-21 Laboratory, Sony Corporation Ph: 81-3-5448-5901. Tokyo, Japan, 1999.

E. Llobet, E.L. Hines, J.W. Gardner and S. Franco. Non-Destructive Banana Ripeness Determination Using a Neural Network-Based Electronic Nose. *Measurement Science & Technology* 10: 538–548, 1999.

D.E. Moriarty, A.C. Schultz and J.J. Grefenstette. Evolutionary Algorithms for Reinforcement Learning. *Journal of Artificial Intelligence Research,* 11: 199–229, 1999.

J.C. Morrison and T.T. Nguyen. On-Board Software for the Mars Pathfinder Microrover. *Jet Propulsion Laboratory report IAA-L-0504P,* 1996.

S. Nolfi and D. Parisi. Growing Neural Networks. *Institute of Psychology, National Research Council Technical Report PCIA-91-15,* Rome, Italy, 1991. (Presented at Artificial Life III, Santa Fe, New Mexico, June 15–19, 1992.)

M.A.L. Oshana. The Misguided Marriage of Responsibility and Autonomy. *The Journal of Ethics,* 6: 261–280, 2002.

OTIS Elevators. Redefining Elevator Performance, Safety and Comfort: The OTIS Elevonic Class. Product Description, 2003. Available: http://www.otis.com

K. Sasaki, S. Markon and M. Makagawa. Elevator Group Supervisory Control System Using Neural Networks. *Elevator World,* 1, 1996.

Schindler Elevator Corporation AITP: Artificial Intelligence Traffic Processor. Technical Product Description, 2003. Available: http://www.us.schindler.com

A.C. Schultz. Using a Genetic Algorithm to Learn Strategies for Collision Avoidance and Local Navigation. In *Proceedings of the Seventh International Symposium on Unmanned Untethered Submersible Technology,* (pp. 213–215). University of New Hampshire Marine Systems Engineering Laboratory, New Hampshire, September 23–25, 1991.

A.C. Schultz. Learning Robot Behaviors Using Genetic Algorithms. In. *Proceedings of the International Symposium on Robotics and Manufacturing*. Washington DC, August 14–18, 1994.

S.B. Stancliff and M.C. Nechyba. Learning to Fly: Modeling Human Control Strategies in an Aerial Vehicle. Machine Intelligence Laboratory, Electrical and Computer Engineering, University of Florida, 2000. Available: http://www.mil.ufl.edu/publications.

W. Stolzmann, M.V. Butz, J. Hoffmann and D.E. Goldberg. First Cognitive Capabilities in the Anticipatory Classifier System. *Illinois Genetic Algorithms Laboratory Report No. 2000008,* University of Illinois, Urbana, 2000.

P. Strawson. Freedom and Resentment. *Proceedings of the British Academy*, 48, 1962.

B.-T. Zhang and Y.-W. Seo. Personalized Web-Document Filtering Using Reinforcement Learning. AI Lab, School of Computer Science and Engineering, Seoul National University, Korea, 2000. Available: http://www.scai.snu.ac.kr

Z.H. Zhou, Y. Jiang, Y.B. Yang and S. F. Chen. Lung Cancer Cell Identification Based on Artificial Neural Network Ensembles. *Artificial Intelligence in Medicine,* 24(1): 25–36, 2002.

Law

Principles of International Law Recognized in the Charter of the Nürnberg Tribunal and in the Judgment of the Tribunal

United Nations, International Law Commission · 1950
https://legal.un.org/ilc/texts/instruments/english/draft_articles/7_1_1950.pdf

Principles of International Law Recognized in the Charter of the Nürnberg Tribunal and in the Judgment of the Tribunal

Principle I

Any person who commits an act which constitutes a crime under international law is responsible therefor and liable to punishment.

Principle II

The fact that internal law does not impose a penalty for an act which constitutes a crime under international law does not relieve the person who committed the act from responsibility under international law.

Principle III

The fact that a person who committed an act which constitutes a crime under international law acted as Head of State or responsible Government official does not relieve him from responsibility under international law.

Principle IV

The fact that a person acted pursuant to order of his Government or of a superior does not relieve him from responsibility under international law, provided a moral choice was in fact possible to him.

Principle V

Any person charged with a crime under international law has the right to a fair trial on the facts and law.

Principle VI

The crimes hereinafter set out are punishable as crimes under international law:

(*a*) Crimes against peace:

(i) Planning, preparation, initiation or waging of a war of aggression or a war in violation of international treaties, agreements or assurances;

(ii) Participation in a common plan or conspiracy for the accomplishment of any of the acts mentioned under (i).

(*b*) War crimes:

Violations of the laws or customs of war which include, but are not limited to, murder, ill-treatment or deportation to slave-labour or for any other purpose of civilian population of or in occupied territory, murder or ill-treatment of prisoners of war, of persons on the seas, killing of hostages, plunder of public or private property, wanton destruction of cities, towns, or villages, or devastation not justified by military necessity.

(*c*) Crimes against humanity:

Murder, extermination, enslavement, deportation and other inhuman acts done against any civilian population, or persecutions on political, racial or religious grounds, when such acts are

done or such persecutions are carried on in execution of or in connection with any crime against peace or any war crime.

Principle VII

Complicity in the commission of a crime against peace, a war crime, or a crime against humanity as set forth in Principle VI is a crime under international law.

_____

Protocol Additional to the Geneva Conventions (Protocol I): Articles 36 and 57

International Committee of the Red Cross, IHL Databases · 8 June 1977
https://ihl-databases.icrc.org/en/ihl-treaties/api-1977

Article 36 – New weapons

In the study, development, acquisition or adoption of a new weapon, means or method of warfare, a High Contracting Party is under an obligation to determine whether its employment would, in some or all circumstances, be prohibited by this Protocol or by any other rule of international law applicable to the High Contracting Party.

Article 57 – Precautions in attack

1. In the conduct of military operations, constant care shall be taken to spare the civilian population, civilians and civilian objects.

2. With respect to attacks, the following precautions shall be taken:

(a) those who plan or decide upon an attack shall:

(i) do everything feasible to verify that the objectives to be attacked are neither civilians nor civilian objects and are not subject to special protection but are military objectives within the meaning of paragraph 2 of Article 52 and that it is not prohibited by the provisions of this Protocol to attack them;

(ii) take all feasible precautions in the choice of means and methods of attack with a view to avoiding, and in any event to minimizing, incidental loss of civilian life, injury to civilians and damage to civilian objects;

(iii) refrain from deciding to launch any attack which may be expected to cause incidental loss of civilian life, injury to civilians, damage to civilian objects, or a combination thereof, which would be excessive in relation to the concrete and direct military advantage anticipated;

(b) an attack shall be cancelled or suspended if it becomes apparent that the objective is not a military one or is subject to special protection or that the attack may be expected to cause incidental loss of civilian life, injury to civilians, damage to civilian objects, or a combination thereof, which would be excessive in relation to the concrete and direct military advantage anticipated;

(c) effective advance warning shall be given of attacks which may affect the civilian population, unless circumstances do not permit.

3. When a choice is possible between several military objectives for obtaining a similar military advantage, the objective to be selected shall be that the attack on which may be expected to cause the least danger to civilian lives and to civilian objects.

4. In the conduct of military operations at sea or in the air, each Party to the conflict shall, in conformity with its rights and duties under the rules of international law applicable in armed conflict, take all reasonable precautions to avoid losses of civilian lives and damage to civilian objects.

5. No provision of this Article may be construed as authorizing any attacks against the civilian population, civilians or civilian objects.

UK GDPR: Article 4(7)–(8) and Article 28 (Processor)

legislation.gov.uk · 2016/679, as retained in UK law

https://www.legislation.gov.uk/eur/2016/679/article/28

Article 4 – Definitions (extract)

(7) ‘controller’ means the natural or legal person, public authority, agency or other body which, alone or jointly with others, determines the purposes and means of the processing of personal data [(but see section 6 of the 2018 Act)] ;

(8) ‘processor’ means a natural or legal person, public authority, agency or other body which processes personal data on behalf of the controller;

Article 28 – Processor

1. Where processing is to be carried out on behalf of a controller, the controller shall use only processors providing sufficient guarantees to implement appropriate technical and organisational measures in such a manner that processing will meet the requirements of this Regulation and ensure the protection of the rights of the data subject.

2. The processor shall not engage another processor without prior specific or general written authorisation of the controller. In the case of general written authorisation, the processor shall inform the controller of any intended changes concerning the addition or replacement of other processors, thereby giving the controller the opportunity to object to such changes.

3. Processing by a processor shall be governed by a contract or other legal act under [domestic law], that is binding on the processor with regard to the controller and that sets out the subject-matter and duration of the processing, the nature and purpose of the processing, the type of personal data and categories of data subjects and the obligations and rights of the controller. That contract or other legal act shall stipulate, in particular, that the processor:

(a) processes the personal data only on documented instructions from the controller, including with regard to transfers of personal data to a third country or an international organisation, unless required to do so by [domestic law] ; in such a case, the processor shall inform the controller of that legal requirement before processing, unless that law prohibits such information on important grounds of public interest;

(b) ensures that persons authorised to process the personal data have committed themselves to confidentiality or are under an appropriate statutory obligation of confidentiality;

(c) takes all measures required pursuant to Article 32;

(d) respects the conditions referred to in paragraphs 2 and 4 for engaging another processor;

(e) taking into account the nature of the processing, assists the controller by appropriate technical and organisational measures, insofar as this is possible, for the fulfilment of the controller's obligation to respond to requests for exercising the data subject's rights laid down in Chapter III;

(f) assists the controller in ensuring compliance with the obligations pursuant to Articles 32 to 36 taking into account the nature of processing and the information available to the processor;

(g) at the choice of the controller, deletes or returns all the personal data to the controller after the end of the provision of services relating to processing, and deletes existing copies unless [domestic law] requires storage of the personal data;

(h) makes available to the controller all information necessary to demonstrate compliance with the obligations laid down in this Article and allow for and contribute to audits, including inspections, conducted by the controller or another auditor mandated by the controller.

With regard to point (h) of the first subparagraph, the processor shall immediately inform the controller if, in its opinion, an instruction infringes this Regulation or [other domestic law relating to data protection].

4. Where a processor engages another processor for carrying out specific processing activities on behalf of the controller, the same data protection obligations as set out in the contract or other legal act between the controller and the processor as referred to in paragraph 3 shall be imposed on that other processor by way of a contract or other legal act under [domestic law], in particular providing sufficient guarantees to implement appropriate technical and organisational measures in such a manner that the processing will meet the requirements of this Regulation. Where that other processor fails to fulfil its data protection obligations, the initial processor shall remain fully liable to the controller for the performance of that other processor's obligations.

5. Adherence of a processor to an approved code of conduct as referred to in Article 40 or an approved certification mechanism as referred to in Article 42 may be used as [a means of demonstrating] sufficient guarantees as referred

to in paragraphs 1 and 4 of this Article.

6. Without prejudice to an individual contract between the controller and the processor, the contract or the other legal act referred to in paragraphs 3 and 4 of this Article may be based, in whole or in part, on standard contractual clauses referred to in [paragraph 8] of this Article, including when they are part of a certification granted to the controller or processor pursuant to Articles 42 and 43. 7.....

8. [The [Commission]] may adopt standard contractual clauses for the matters referred to in paragraph 3 and 4 of this Article....

9. The contract or the other legal act referred to in paragraphs 3 and 4 shall be in writing, including in electronic form.

10. Without prejudice to Articles 82, 83 and 84, if a processor infringes this Regulation by determining the purposes and means of processing, the processor shall be considered to be a controller in respect of that processing.