Peace of Mind: Cognitive Warfare and the Governance of Subversion in the 21st Century

Jean-Marc Rickli, Federico Mantellassi
and Gwyn Glasser

August 2023

GCSP Policy Brief No.9

GCSP
Geneva Centre for
Security Policy

Geneva Centre for Security Policy

The Geneva Centre for Security Policy (GCSP) is an international foundation serving a global community of organisations and individuals. The Centre's mission is to advance peace, security and international cooperation by providing the knowledge, skills and network for effective and inclusive decision-making through executive education, diplomatic dialogue, research and policy advice.

The GCSP Policy Briefs Series

The GCSP Policy Briefs series addresses current security issues, deduces policy implications and proposes policy recommendations. It aims to directly inform policy- and decision-making of states, international organisations and the private sector.

Under the leadership of Ambassador Thomas Greminger, Director of the GCSP, the series is edited by Professor Nayef Al-Rodhan, Head of the Geopolitics and Global Futures Programme, and Mr Tobias Vestner, Head of the Research and Policy Advice Department, and managed by Ms Christine Garnier Simon, Administration and Coordination Officer, GCSP Geopolitics and Global Futures.

Geneva Centre for Security Policy
Maison de la paix
Chemin Eugène-Rigot 2D
P.O. Box 1295
1211 Geneva 1
Switzerland
Tel: + 41 22 730 96 00
E-mail: info@gcsp.ch
www.gcsp.ch

ISBN: 978-2-88947-416-5

About the authors

Dr Jean-Marc Rickli is the Head of Global and Emerging Risks and the Founder and Director of the Polymath Initiative at the GCSP. He is also the co-chair of the NATO Partnership for Peace Consortium (PfPC) Emerging Security Challenges Working Group and a senior advisor for the Artificial Intelligence Initiative at the Future Society. He is the co-curator of the International Security Map of the Strategic Intelligence Platform of the World Economic Forum. He is also a member of the Geneva University Committee for Ethical Research and of the advisory board of Tech4Trust, the first Swiss startup acceleration program in the field of digital trust and cybersecurity. Prior to these appointments, Dr Rickli was an assistant professor at the Department of Defence Studies of King's College London and at the Institute for International and Civil Security at Khalifa University in Abu Dhabi. In 2020, he was nominated as one of the 100 most influential French-speaking Swiss by the Swiss newspaper *Le Temps*. Dr Rickli received his PhD in International Relations from Oxford University. His latest book published by Georgetown University is entitled *Surrogate Warfare: The Transformation of War in the Twenty-first Century*.

Mr Federico Mantellassi is a Research and Project Officer at the Geneva Centre for Security Policy where he has worked since 2018. Federico's research and writing focuses on how emerging technologies impact international security and warfare, as well as on the societal implications of their development and use. Federico is also the project coordinator of the GCSP's Polymath Initiative; an effort to create a community of scientists able bridge the gap between the scientific and technological community and the world of policy making. Previously, he assisted in the organisation of executive education activities at the GCSP and was the project coordinator of the annual Geneva Cyber 9/12 Strategy Challenge. He holds a Master's Degree in Intelligence and International Security from King's College London, and a Bachelor's Degree in International Studies from the University of Leiden. Federico speaks English, French and Italian.

Mr Gwyn Glasser is a consultant on technology governance for the Cyberpeace institute, and a former Junior Researcher at the Geneva Center for Security Policy where he focused on governance of strategic operations enabled by AI and emerging technologies, including cognitive warfare and lethal autonomous weapons. He was also a consultant for INHR on governance of AI military systems and worked with NGOs ForHumanity and All Tech is Human, contributing to the discourse on AI and human rights, and training on applied algorithm ethics. Gwyn is an AiCore certified machine learning engineer and holds a MA in Philosophy from the University of Edinburgh, specializing in AI Ethics. He speaks English, French, Italian and Japanese.

Introduction

The 21st century is marked by unprecedented and exponential technological advances.[1] These advances are changing the ways in which states influence, coerce, subvert and wage war, and have democratised access to global means of influence for non-state actors – and even individuals.[2] This has accelerated the prevalence of so-called hybrid forms of warfare. An "age of unpeace" appears to have established itself, with a pervasive sense of both non-peace and non-war.[3] Influence campaigns utilising digitally spread disinformation and instrumentalising social media technologies have proliferated, enhancing ways to target the human mind. Current and future developments in artificial intelligence (AI), cognitive sciences, neurotechnologies, and other related fields will further increase the risks of mass manipulation and lead to the possibility of the militarisation of the mind as the battlefield of the future.

The emergence of the *cognitive domain* as the sixth domain of warfare[4] will lead to the increased conduct of *cognitive warfare*, which is likely to raise the profile and efficiency of non-kinetic means of subversion over kinetic means of coercion, while benefitting from a lack of international governance. If international governance remains static while these tools advance at great speeds, the international system will lack the frameworks, tools, and understanding needed to govern the means of 21st century subversion.

This policy brief explores the emergence of cognitive warfare – which aims at controlling what and how an adversary thinks – and the rise on the international stage of non-kinetic means of subversion enabled by emerging technologies. It proposes the establishment of governance frameworks to regulate the use of emerging technologies for purposes of cognitive warfare with the ultimate aim of subversion. It promotes the concept of "subversion control" in order to prevent the militarisation of the mind. It further recommends the regulation of enabling emerging technologies, such as neurotechnologies and AI, while promoting the enhancement of "societal resilience", especially in democracies, whose open and digitalised information environments make them structurally more vulnerable to the practices of cognitive warfare. Lastly, the brief seeks to promote more research into the concept of cognitive warfare itself, which will in turn assist governance efforts.

[1] A. Azeem, *Exponential: Order and Chaos in an Age of Accelerating Technology*, New York, Penguin Random House, 2021.

[2] A. Krieg and J.-M. Rickli, *Surrogate Warfare: The Transformation of War in the Twenty First Century*, Washington DC, Georgetown University Press, 2019.

[3] M. Leonard, *The Age of Unpeace*, London, Penguin Random House, 2021.

[4] The current domains of warfare include land, sea, air, outer space and cyberspace; see "Cognitive warfare", below.

The security challenge

Subversion

Subversion is at the cusp of a paradigmatic change in its use on the world stage by various actors. While subversive tactics themselves are not new, recent emerging technologies are enabling subversion at an unprecedented scale, far greater granularity and increased accessibility by state and non-state actors. Cognitive warfare – or the sum of cognitive operations aiming at the control of the adversary's mind, perceptions and action – presents an increasingly viable alternative to using force or diplomacy to achieve strategic objectives. However, the cognitive domain of warfare is subject to little or no international governance. There is thus a need for the international community to build a sophisticated understanding of the cognitive domain and cognitive warfare, their role in enabling more efficient subversion, and of potential governance frameworks for operations within this space.

Subversion can be defined as an "instrument of power used in non-military covert operations. It exploits vulnerabilities to secretly infiltrate a system of rules and practices in order to control, manipulate, and use the system to produce detrimental effects against an adversary".[5] Typical subversion mechanisms have relied on spies or other means to infiltrate and influence adversaries' institutions turning social or other systems against the adversary with a range of possible context-sensitive mechanisms and effects. These include "influence on public opinion, disintegration of social cohesion, economic disruption, infrastructure sabotage, influence on government policy, and, in the extreme case, overthrowing a government".[6] However, the effectiveness of subversion has been limited by various factors such as:

1. *Resources:* Traditional subversion tactics, particularly the use of espionage, require significant resources for preparing agents and infiltrating them into an adversary's institutions.[7]
2. *Coordination:* Achieving a strategically significant impact from subversion operations requires a huge organisational capacity, particularly in long-term erosion operations that aim to achieve an objective by a culmination of diverse activities.[8]

[5] L. Maschmeyer, "Subversive Trilemma: Why Cyber Operations Fall Short of Expectations", *International Security*, Vol.46(2), 2021a, pp.51-90, https://doi.org/10.1162/isec_a_00418.

[6] Ibid.

[7] Ibid.

[8] Ibid.

3. *Scope:* Subversion must target a sufficiently large audience to achieve a strategic impact when its objective is to shift public opinion at the national level.[9]

4. *Granularity:* Fundamentally, subversion involves accessing and manipulating the minds of individuals. A subversion operation must be able to operate at a sufficiently granular level to interface with the complexities of individual thought patterns and the unique dynamics and vulnerabilities of relevant groups or institutions.[10]

Due to these limiting factors, effective subversion operations have generally been carried out by powerful states with the necessary capacities, but usually with limited impact. Russia's inability to achieve its strategic goals through its long-running subversive operations in Ukraine (cyber-attacks, disinformation campaigns) is one such example.[11] Today, however, developments in the field of neurotechnology, cognitive sciences, and AI and the democratisation of related technologies fundamentally alter these restricting factors. By enabling more accurate, larger scale, and less costly cognitive operations, subversion will become increasingly accessible at much lower cost, with mechanisms for globally impactful operations at high levels of granularity and with automated coordination. This strongly points to a near future in which subversion remains understudied and ungoverned, while manifesting exponentially increasing potential for influencing and manipulating adversaries, and with negligible barriers for entry for both state and non-state actors.

Cognitive warfare

Emerging technologies such as AI (especially generative AI) or neurotechnologies are enabling highly accessible and efficient subversion within the cognitive domain of warfare.[12] Warfare is understood as unfolding within – and across – domains, commonly defined as the different operational environments in which military operations take place. These have traditionally been the geographical spaces where contact with the enemy occurs, namely land, sea, air, outer space and, more recently, cyberspace.[13] The domains of war

[9] Ibid.

[10] Ibid.; L. Maschmeyer, "Subversion, Cyber Operations, and Reverse Structural Power in World Politics", *European Journal of International Relations*, Vol.29(1), 2022a, https://doi.org/10.1177/13540661221117051.

[11] L. Maschmeyer et al., "Donestk Don't Tell – 'Hybrid War' in Ukraine and the Limits of Social Media Influence Operations", *Journal of Information Technology and Politics*, 2023, doi:10.1080/19331681.2023.2211969; L. Maschmeyer and M. Dunn Cavelty, "Goodbye Cyberwar: Ukraine as Reality Check", *ETH Zurich Centre for Security Policy*, Vol.10(3), 2022, https://ethz.ch/content/dam/ethz/special-interest/gess/cis/center-for-securities-studies/pdfs/PP10-3_2022-EN.pdf.

[12] J.-M Rickli. "Neurotechnologies and Future Warfares," *RSIS, Nanyang Technological University*, 7 December 2020, https://www.rsis.edu.sg/rsis-publication/rsis/ai-governance-and-military-affairs-neurotechnologies-and-future-warfare/#.YAp-Oi2ZPEZ

[13] C. McGuffin and P. Mitchell, "On Domains: Cyber and the Practice of Warfare", *International Journal*, Vol.69(3), 2014, pp.394-412, 10.1177/0020702014540618.

have long been tightly linked to technological innovation, with advances in naval and aviation technologies, for example, opening up the seas and skies to warfare.[14] Most recently, and perhaps more controversially, digital technologies have opened up the cyber domain. Aside from the domain in which they occur, activities in warfare can be further divided between kinetic actions (i.e. actions that have a physical effect) and non-kinetic actions. While "hot" conflicts continue to rage across the world, most obviously in Ukraine, confrontations between major military powers have since the end of the Second World War occurred mostly either through proxies and surrogates, often below the threshold of war, or increasingly often through non-kinetic means.

Most commonly, this type of warfare has been termed "hybrid warfare", which can be characterised as:

> a creative act of force combining a broad spectrum of military and non-military instruments and vectors of power on an extended multi-domain battlespace ... while ambiguously operating in the shadow/grey-zones of blurred interfaces – between war and peace, friend and foe, internal and external relations, civil and military as well as state and non-state actors and fields of responsibilities – with the ultimate goal to enable an own decision of the confrontation primarily on non-military centres of gravity while preventing being militarily overthrown or compelled by the enemy.[15]

The prevalence of hybrid forms of warfare in the 21st century – at least between great powers – has blurred the lines between domains, the foreign and domestic spheres, state and non-state actors, and peace and war. The resulting environment is one of "permanent latent struggles"[16] rather than a clearly delineated state of peace and war. This state has been referred to as "new generational warfare",[17] "unpeace"[18] or conflict in the "noosphere".[19]

---

[14] Ibid.

[15] J. Schmid, "Introduction to Hybrid Warfare – A Framework for Comprehensive Analysis", in R. Thiele (ed.), *Hybrid Warfare: Future and Technologies*, London, Routledge, 2021, pp.11-33, https://doi.org/10.1007/978-3-658-35109-0.

[16] D. Wurm, "Cognitive Domain: Hybrid and Future Forms of Prosecuting Conflicts in the Cognitive Domain", Austrian Federal Ministry of Defence, 2022.

[17] D. Pappalardo, "Win the War before the War?: A French Perspective on Cognitive Warfare", War on the Rocks, 1 August 2022, https://warontherocks.com/2022/08/win-the-war-before-the-war-a-french-perspective-on-cognitive-warfare/.

[18] Maschmeyer, 2021a.

[19] B. Claverie and F. Du Cluzel, "Cognitive Warfare: The Advent of the Concept of 'Cognitics' in the Field of Warfare", in B. Claverie et al. (eds), *Cognitive Warfare: The Future of Cognitive Dominance*, HAL Open Science, April 2022, https://hal.science/hal-03635889/document#:~:text=Cognitive%20warfare%20is%20thus%20an,the%20individual%20and%20collective%20levels.

Recent scholarship is now beginning to emerge focusing on the concept of cognitive warfare.[20] While no consensus exists on the definition of the term, the following working definition highlights key characteristics from discussions across the literature: *cognitive warfare is any subversion operation aimed at affecting the mechanisms of understanding and decision-making [21] of individuals and/or populations,[22] in order to achieve strategic objectives*. Cognitive operations "can be used before, during and after kinetic actions, while remaining outside current international definitions of what constitutes an act of war".[23] These operations will also become far more prevalent and effective with the increasingly rapid development of emerging technologies such as AI or neurotechnologies, and with the increasing integration of these technologies into daily life. Cognitive warfare is distinct from information warfare in that information warfare "focuses on controlling the flow of information", while cognitive warfare "aims to control the responses of targets to the presented information".[24] As Hung and Hung note, "Although ... cyberwarfare, information warfare, cognitive warfare, and hybrid warfare ... contain the element of influence operations and may impact human cognition, only cognitive warfare is specifically dedicated to brain control by incorporating weaponized neuroscience into various practices". [25] Cognitive warfare targets a nation's entire human capital; humans – and their cognitive space – become the contested domain.

Cognitive warfare is increasingly effective at meeting the objectives of subversion – i.e. manipulation, disruption or overthrowing governments. One of the most prominent examples of the impact of cognitive operations is the interference with the 2016 US presidential elections. Russia's Internet Research Agency was responsible for the creation of fake social media accounts across all available platforms, garnering over 263.5 million active engagements with their content on Facebook and Instagram alone,[26] with at least 50,000 bot accounts and over 3,000 fake accounts regularly posting on Twitter in the period leading up to the election.[27] These interventions

[20] F. Du Cluzel, "Cognitive Warfare", NATO Innovation Hub, November 2020, https://www.innovationhub-act.org/sites/default/files/2021-01/20210113_CW%20Final%20v2%20.pdf.

[21] Pappalardo, 2022.

[22] Du Cluzel, 2020.

[23] Claverie and Du Cluzel, 2022.

[24] T.C. Hung and T.W. Hung, "How China's Cognitive Warfare Works: A Frontline Perspective on Taiwan's Anti-Disinformation Wars", *Journal of Global Security Studies*, Vol.7(4), 2022, https://academic.oup.com/jogss/article/7/4/ogac016/6647447.

[25] Ibid.

[26] S. Shane and S. Frenkel, "Russian 2016 Influence Operation Targeted African-Americans on Social Media", *New York Times*, 17 December 2018, https://www.nytimes.com/2018/12/17/us/politics/russia-2016-influence-campaign.html.

[27] J. Swaine, "Twitter Admits Far More Russian Bots Posted on Election than It Had Disclosed", *The Guardian*, 20 January 2018, https://www.theguardian.com/technology/2018/jan/19/twitter-admits-far-more-russian-bots-posted-on-election-than-it-had-disclosed.

took a range of positions, including encouraging African-American, Mexican-American, and Hispanic voters to boycott or mistrust elections; empowering far right-wing voters; and spreading disinformation to voters on both sides of the political spectrum, thus contributing to the polarisation of the US domestic population and the disruption of the elections. [28] Cognitive operations are being deployed in a range of other arenas, such as Russia's campaigns "to diminish trust in and within democratic nations globally",[29] and by non-state actors such as Al-Qaeda or Islamic State in promoting radicalisation and supporting their own recruitment campaigns.[30]

While the efficacy of these operations is difficult to quantify, this lack of clarity around what constitutes cognitive attacks, their impacts and how to respond to them is a strong motivator for their use. The indirect impacts of cognitive warfare, as well as its tendency to co-opt independent agents by affecting their beliefs, offers the benefits of plausible deniability of responsibility by those carrying out such attacks and even sometimes legal protections in democratic states that value free speech. [31] Thus, cognitive attacks can currently be deployed with little risk of reprisals from the international community, and are particularly effective against democratic states because of these states' openness.

These factors are compounded by highly accessible emerging technologies that will continue to boost the efficiency of cognitive warfare. These include advances in cognitive technologies, such as technologies that monitor and interface with the brain (such as brain-computer interface (BCI) technology) or that replicate human cognition (such as machine learning). Furthermore, the rapid merging of social and technological systems expands the target area for cognitive attacks. Thus, emerging technologies enhance the potential impact of subversion through cognitive attacks in two ways:

1. *Indirectly:* The integration of information technology into daily life creates new and highly frequented online spaces that can be targeted (social media such as Twitter, TikTok, Instagram, Facebook, metaverses or future spaces accessible via BCIs), and generates more and more granular user data for manipulation that can be used to resource cognitive attacks.[32]

[28] A. Bernal et al., "Cognitive Warfare: An Attack on Truth and Thought", NATO Innovation Hub and Johns Hopkins University, 2020, https://www.innovationhub-act.org/sites/default/files/2021-03/Cognitive%20Warfare.pdf.

[29] Ibid.

[30] Ibid.

[31] Ibid.

[32] J.-M. Rickli and F. Mantellassi, "Our Digital Future: The Security Implications of Metaverses", Strategic Security Analysis No. 24, Geneva Centre for Security Policy, March 2022, https://dam.gcsp.ch/files/doc/ssa-25-march-2022.

2. *Directly:* by developing tools that can be used in cognitive operations, such as those that coordinate complex widespread operations, profile individuals' behaviour, and/or generate targeted content, and even those that directly interact with people's brains. Examples include AI-coordinated bot networks, AI-generated content, false sense perceptions caused by BCIs, or "emotional AI" that adapts according to the target's emotional state.[33]

These advances bypass the traditional limitations of subversion operations. The mass production of data and automated content creation lead to a publicly available abundance of data that can be used for cognitive manipulations. Algorithms can now coordinate accurate profiling and targeting across a massive population, with a granular approach that adapts strategies according to specific individual behaviours.[34] Large Language Models (LLMs) such as ChatGPT could be used to collect data so as to "understand" their interlocutor and to nudge and influence with great efficiency by knowing through the use of which prompts, methods and topics and at which times of day an interlocutor is most susceptible to manipulation, while updating this data in real time.[35] The impact of such models will be enhanced as their access to the full range of an individual's activities, thoughts and behaviours increases as populations integrate increasingly deeper into digital spaces. The advent of non-invasive brain monitoring technologies such as electroencephalography (EEG), magnetoencephalography (MEG) or functional magnetic resonance imaging (fMRI) and invasive technologies such as BCIs further exacerbates these vulnerabilities, allowing the recording of the neural processes of the targeted subject and even directly influencing their thoughts patterns.[36]

Subversion has long existed as an integral part of warfare and geopolitics. However, today's technological advances represent both a quantitative shift due to the proliferation of neuro-affecting devices and tools and a qualitative shift due to the efficiency and utility of emerging technologies, which provide unprecedented ability to affect the cognitive domain. It follows that cognitive warfare will benefit from these developments and accelerate the use of subversion to exert power globally. Means of influence will gradually shift away from purely kinetic approaches towards subversion.

[33] V. Bakir and A. McStay, "Fake News and the Economy of Emotions", *Digital Journalism*, Vol.6(2), 2017, pp.154-175, https://doi.org/10.1080/21670811.2017.1345645.

[34] S. Zuboff, *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*, London, Profile Books, 2019; T. Christiano, "Algorithms, Manipulation, and Democracy", *Canadian Journal of Philosophy*, Vol.52(1), 2022, pp.109-124, https://doi.org/10.1017/can.2021.29.

[35] N. Sanders and B. Schneider, "How ChatGPT Hijacks Democracy", *New York Times*, 15 January 2023, https://www.nytimes.com/2023/01/15/opinion/ai-chatgpt-lobbying-democracy.html; G. Marcus, "Why Are We Letting the AI Crisis Just Happen?" *The Atlantic*, 13 March 2023, https://www.theatlantic.com/technology/archive/2023/03/ai-chatbots-large-language-model-misinformation/673376

[36] Rickli and Mantellassi, 2022.

Political ends may be much more effectively achieved through targeted, automated cognitive operations than by using force or threatening to do so. No governance framework or working tools currently exist to deal with the dynamics of this state of affairs.[37]

Policy implications

A lack of governance in the cognitive domain creates a legal, normative, and ethical vacuum around the new tools state and non-state actors can utilise to wage a new kind of war. This represents a substantial risk and a departure from the now-established practice of seeking to reduce the likelihood and effects of war. It could lead to the erosion of ethical and legal norms governing warfare, the unrestricted militarisation of the mind and violation of mental privacy, and technological arms races in the field of neurotechnologies and neuroweapons, and thus to an overall decrease in international stability.

An ungoverned space

International governance frameworks such as international humanitarian law (IHL) and various arms control agreements exist to increase stability, reduce the likelihood of war, and lessen suffering in the event of its occurrence.[38] The very basis of these frameworks – such as what constitutes an "attack" or a "weapon" – is inextricably linked to the use of physical force, territoriality and recognisable effects (such as visible injuries or destruction).[39] However, these ways of conceptualising aggression cannot be transposed to the realities of cognitive warfare.

As we have demonstrated, cognitive warfare will increase the pace and efficiency of subversion – which will largely be non-kinetic in nature – as a tool of power. Already, the current focus on the kinetic impact of warfare has created an interpretive gap in IHL regarding non-kinetic acts that is leveraged by hybrid warfare.[40] As cognitive warfare and the "weapons" used to wage it will not fit into the way we conceptualise warfare in order to govern it, traditional governance frameworks will no longer be fit for purpose.[41] Hybrid tactics such as disinformation campaigns or cyber-attacks

[37] R. Chavarriaga et al., "Neurotechnologies: The New Frontier for International Governance", Strategic Security Analysis No. 29, Geneva Centre for Security Policy, April 2023, https://dam.gcsp.ch/files/doc/ssa-2023-issue29.

[38] J. Mauer, "The Purposes of Arms Control", *Texas National Security Review*, Vol.2(1), 2018, http://dx.doi.org/10.26153/tsw/870; ICRC (International Committee of the Red Cross),"What Is International Humanitarian Law?", ICRC, 2022, https://www.icrc.org/en/document/what-international-humanitarian-law.

[39] Wurm, 2022.

[40] Ibid.

[41] J.-M. Rickli, "Does the UN Need a Watchdog to fight Deepfakes and Other AI Threats," World Economic Forum, 2 August, 2023, https://www.weforum.org/agenda/2023/08/does-un-needs-watchdog-fight-

have been destabilising and subversive, but have so far had a mixed track record of achieving strategic ends.[42] The impact of a lack of governance has therefore not been catastrophic thus far. However, as cognitive warfare turns subversion into an increasingly effective tool to achieve such ends, the implications of the lack of an internationally agreed governance regime become more consequential.

Governance efforts focused on the mind are already starting to emerge, most outwardly in the concept of neurorights.[43] Championed by the Neurorights Foundation, neurorights seek to enhance the human rights framework, creating a new category of rights to more appropriately protect the human mind.[44] Parallel efforts should therefore aim to govern the tools – neuroweapons and affiliated means – and ways through which these rights might be threatened. Today's governance regimes, which are principally concerned with the use of kinetic force, need to be updated to reflect the non-kinetic capabilities that enable cognitive warfare. As such, rules of responsible state behaviour; limits on the types of "weapons" and how they can be utilised; and legal, ethical, and moral guidelines should be developed to this effect.

International stability

Without appropriate governance regimes, incentives to develop, militarise and deploy relevant capabilities to gain strategic advantages could multiply. Because emerging technologies such as AI and neurotechnologies are key enablers of cognitive warfare, states will likely feel the need to militarise these dual-use technologies. This could in turn accelerate technological arms races in these fields. Arms race dynamics could induce a race to the bottom in terms of ethical, legal, and normative restraints on the development and use of these technologies. Because states will be fearful of falling behind, they are likely to forgo strict regulation so as to benefit from these technologies' most disruptive capabilities for matters of national security.[45] This could normalise the militarisation of the mind and lead to unrestricted irresponsible behaviour in the cognitive domain. Ultimately,

deepfakes-ai-threats/

[42] L. Maschmeyer, "Digital Disinformation: Evidence from Ukraine", CCS Analyses in Security Policy No. 278, ETH Zurich, February 2021b, https://css.ethz.ch/content/dam/ethz/special-interest/gess/cis/center-for-securities-studies/pdfs/CSSAnalyse278-EN.pdf; L. Maschmeyer, "Subversion over Offence: Why the Practice of Cyber Conflict Looks Nothing Like Its Theory and What This Means for Strategy and Scholarship", Offensive Cyber Working Group, 2022b, https://offensivecyber.org/2022/01/19/subversion-over-offense-why-the-practice-of-cyber-conflict-looks-nothing-like-its-theory-and-what-this-means-for-strategy-and-scholarship/.

[43] M. Ienca et al., "Towards a Governance Framework for Brain Data", *Neuroethics*, 2022, https://link.springer.com/article/10.1007/s12152-022-09498-8.

[44] Chavarriaga et al., 2023.

[45] A. Holland Michel, "Recalibrating Assumption on AI", Chatham House, 12 April 2023, https://www.chathamhouse.org/2023/04/recalibrating-assumptions-ai/04-race-no-winners.

these dynamics could prevent the emergence of norms around mental privacy, manipulation, influence, self-determination and integrity. As these technologies become key enablers of subversion and vital national security tools, the already apparent trend of great-power technological decoupling could also accelerate.

Cognitive warfare could also have wholesale consequences for the offence-defence balance. Proponents of Offence-Defence Balance Theory hold that international stability can be linked to whether we live in an environment in which offence or defence has the advantage, with the former prompting instability and conflict and the latter stability and peace.[46] They further argue that the main variable affecting this balance is technology, which can tilt the balance either way.[47] In this respect, the ability to wage effective cognitive warfare may create escalatory pressures and favour the offensive use of these capabilities.

Indeed, there is inherent difficulty in detecting when a cognitive attack is taking place and in the ability to defend oneself from it. The possibility of being under cognitive attack at any given moment could incentivise "first use" dynamics in conducting cognitive warfare. Plausible deniability is a key incentive in the use of surrogates in warfare, including technological surrogates.[48] In the case of cognitive warfare, the undefined and unregulated nature of the tools used to wage it, as well as their efficiency, could fuel such plausible deniability and further incentivise first use. Additionally, cognitive warfare will become even more destabilising as its enabling technologies democratise and proliferate, and the ability to influence huge numbers of people becomes accessible to non-state actors, corporations, and even individuals.

Policy recommendations

From the security challenges and policy implications described above, the following policy recommendations can be deduced:

- The international community should work to devise an international governance framework for "subversion control". The international community has devised tools such as IHL and arms control regimes to govern the use of force by limiting the types of weapons available to states, establishing conditions for their use, and generally limiting the ways in which states can coerce each other. Similarly, rules of responsible behaviour, red lines, bans, and soft laws should be developed around

[46] C. Glaser and C. Kaufmann, "What Is the Offence-Defence Balance and How Can We Measure It?", *International Security*, Vol.22(4), 1998, pp.42-82, https://direct.mit.edu/isec/article-abstract/22/4/44/11593/What-Is-the-Offense-Defense-Balance-and-How-Can-We?redirectedFrom=fulltext&utm_source=adwords&utm_campaign=FY22_Instl_ISEC_Search&utm_medium=ppc.

[47] K. Lieber, "Grasping the Technological Peace: The Offence-Defence Balance and International Security", *International Security*, Vol.25(1), 2000, pp.71-104, https://www.jstor.org/stable/2626774.

[48] Krieg and Rickli, 2019.

subversion and its enabling technologies in order to avoid technological arms races, unrestricted technologically enabled influence campaigns, and cognitive warfare. Efforts to govern subversion will better reflect new realities and more effectively control these new weapons and ways to wage war.

- States should work to build societal resilience. In the absence of a current framework to govern subversion achieved through cognitive warfare, states – democracies especially – need to build domestic resilience against cognitive manipulations through a "whole-of-society approach".[49] These new subversion tools and tactics do not aim to cripple an enemy's traditional critical infrastructure, physical assets or military power as such, but are mainly aimed at a population's way of thinking and acting. Therefore, states must be able to withstand continuous, simultaneous and complex hybrid threats directed at their populations. In this respect, sensitising populations to the vulnerabilities of digital information ecosystems and the role that technology plays in creating these vulnerabilities is paramount and should start in basic school education.
- States should deepen efforts aimed at the international regulation of enabling technologies, especially AI and neurotechnologies. Efforts to this end are already under way, such as the European Union's AI act or Chile's national brain data regulation.[50] Until rules are drawn up to govern state behaviour in the employment of such technologies with subversive intent through cognitive warfare, international technology regulation can at least steer the development and deployment of such technologies in ways that minimise their potential negative impacts. In light of the growing convergence between these technologies, increased cross-fertilisation among governance efforts should take place to better prepare for future technological realities.[51]
- More research into cognitive warfare needs to be conducted and promoted. As a new domain of warfare that significantly departs from other traditional domains, cognitive warfare requires additional definitional work. Using the same terminology as for the physical domains of war has already hampered our understanding and subsequent governance of non-physical domains such as the cyber domain.[52] There is an urgent need to avoid recreating interpretive gaps due to an inability to reconcile the terminology we use to govern warfare with the realities of the cognitive

[49] N. Jackson, "Deterrence, Resilience and Hybrid Wars: The Case of Canada and NATO", *Journal of Military and Strategic Studies*, Vol.19(4), 2019, https://jmss.org/article/view/68870/53337.

[50] L. Guzman, "Chile: Pioneering the Protection of Neurorights", *UNESCO Courier*, 2022, https://en.unesco.org/courier/2022-1/chile-pioneering-protection-neurorights; L. Bertuzzi, "Europe's Rulebook for Artificial Intelligence Takes Shape", International Association of Privacy Professionals, 23 May 2023, https://iapp.org/news/a/europes-rulebook-for-artificial-intelligence-takes-shape/.

[51] Chavarriaga et al., 2023.

[52] Wurm, 2022.

domain. Therefore, further research into the cognitive domain of warfare should be aimed at developing the right language to describe activity in that domain. This includes defining thresholds for what constitutes an "attack," "weapon" or "injury" in the cognitive domain. This might entail shifting paradigms away from solely physical and coercion-based understandings of warfare to one that better incorporates non-physical and subversive activities.

- Technological monitoring, foresight approaches and polymath thinking should be promoted. The pace of developments in the digital domains is exponential. Recent examples such as deepfakes and generative AI have demonstrated that governments are often caught off-guard by such developments. In order to mitigate this, resources should be invested in monitoring technological developments that affect the cognitive space and in anticipating the impacts of emerging technologies. To that end, foresight approaches should become more commonplace in policymaking. In this regard, the identification of "weak signals" is critical. This requires "polymath skills", as opposed to the hyper-specialisation that often results in siloed thinking. Polymath thinkers are able to understand technologies and the social, political, economic, normative and strategic environment in which they are developed. These are crucial skills for developing effective governance systems.[53]

[53] An example of the impact of these skillsets can be found in the activities of the GCSP's Polymath Initiative. For more information on this initiative, see: https://www.gcsp.ch/the-polymath-initiative.

Conclusion

As cognitive warfare enables subversion to become an increasingly effective alternative to the use of force, the international community might find itself in a legal, normative and ethical vacuum. Established tools for the governance of the international use of force such as IHL or various arms control agreements will probably no longer provide effective ways to curtail state behaviour. In parallel to the more traditional tools used to govern the use of force, creating new international governance frameworks focused on subversion and cognitive warfare will create a comprehensive governance toolbox that holistically encompasses the capabilities that state and non-state actors can use to coerce and subvert in the 21st century.

By leaving this space ungoverned, technological arms races in AI and neurotechnologies will accelerate, leading to the unrestricted development, militarisation, and deployment of their most disruptive capabilities. Incentives for first use in conducting cognitive operations will probably shift the offence-defence balance towards the offence, leading to unrestricted subversive state behaviour in the cognitive domain and the erosion of emerging norms around mental privacy, integrity and influence. The dual-use nature of the enabling technologies, as well as and their democratisation will lead to an exponential increase in the number of actors that will have the ability to carry out large-scale cognitive manipulation.

The policy recommendations outlined in this policy brief seek to pre-empt this situation by proposing the idea of “subversion control” regimes, regulation of cognitive-warfare-enabling technologies such as AI and neurotechnology; the building of societal resilience to prevent cognitive manipulation; the promotion of research into the cognitive domain of warfare; and the development of anticipation and detection methods such as technological monitoring, foresight approaches, and polymath thinking.

People make peace and security possible

Geneva Centre for Security Policy
Maison de la paix
Chemin Eugène-Rigot 2D
P.O. Box 1295
1211 Geneva 1
Switzerland
Tel: + 41 22 730 96 00
e-mail: info@gcsp.ch
www.gcsp.ch

ISBN: 978-2-88947-416-5

GCSP
Geneva Centre for Security Policy