TA Tomorrow's Affairs
Analysis of today
Assessment of tomorrow

By: TA Analysis

The US and China are seeking rules for a war that no one ordered

US Treasury Secretary Scott Bessent confirmed on 16 September that he will discuss with Chinese Vice-Premier He Lifeng the risks posed by the use of artificial intelligence to both countries.

The meeting is scheduled for this weekend, ahead of the expected meeting between Donald Trump and Xi Jinping on 24 September in Washington. Bessent announced talks on preventing the fragmentation of American and Chinese technological systems, as well as on the security of models whose core parameters are publicly available and models that companies keep closed.

Until now, Washington has mainly conducted its dialogue with Beijing on artificial intelligence through disputes over chip exports, access to computing capacity and the protection of American technology.

Bessent's announcement broadens the agenda to include risks that both countries have an interest in limiting.

On the eve of new interstate talks, a detailed proposal for rules has emerged, intended to prevent a malfunction in an autonomous system, an unauthorised operation, or a misinterpreted cyber-attack from triggering military escalation between the two nuclear powers.

Talks begin ahead of Trump–Xi meeting

The recommendations were published on 9 September by Melanie Sisson, an expert at the Brookings Institution, and Tianjiao Jiang, a professor at Fudan University.

Both participate in the Dialogue on Artificial Intelligence and National Security, which Brookings and Tsinghua University's Center for International Security and Strategy have run since 2019.

This unofficial expert channel seeks to identify issues the two governments can still discuss even when political relations between them are poor.

Sisson and Jiang do not have a mandate to negotiate on behalf of Washington or Beijing, so their document does not bind either government.

It was published ahead of the announced intergovernmental dialogue on artificial intelligence and examines a danger that technical measures alone cannot resolve.

> Increasingly autonomous systems can cause military incidents whose origin and intent cannot quickly be determined by the party under attack

Increasingly autonomous systems can cause military incidents whose origin and intent cannot quickly be determined by the party under attack.

In a crisis between the United States and China, the time available for verification could be measured in minutes.

The proposal includes a ban on AI making autonomous decisions about the use of nuclear weapons, mandatory human approval of cyber-attacks on nuclear command networks, and a dedicated communications link for incidents caused by autonomous systems.

The authors also call for a shared definition of human control, because the same political formulation in two different military systems can conceal completely different degrees of machine autonomy.

The Lima Declaration covers the latter decision

In November 2024 in Lima, Joe Biden and Xi Jinping confirmed that the decision to use nuclear weapons must remain in human hands.

It was the first publicly confirmed joint position of the two states on the role of artificial intelligence in nuclear decision-making.

The agreement, however, covered only the final decision and did not specify what form human control should take in systems that collect data, assess threats, propose targets and recommend a military response.

> The launch order comes at the end of a long process

The launch order comes at the end of a long process. It is preceded by early-warning satellites and radars, intelligence programmes, communication networks, command centres and procedures for verifying whether an attack is real.

AI can be introduced into each of these components, while formal decision-making rights still rest with the president or authorised commander.

A data analysis programme can mislabel routine military movements as preparations for an attack.

The system responsible for defending the network can automatically terminate the connection with the command centre or initiate a countermeasure against the source of the suspicious activity.

An autonomous cyber tool previously inserted into an adversary's infrastructure to collect data may, due to an error, compromise or misassigned task, carry out an operation not authorised by its creators.

A country affected by such a disturbance experiences the consequences long before it knows who caused them.

Losing communications with nuclear units may appear to be the start of an attempt to disarm the adversary before a first strike.

An automatic defensive reaction can be interpreted as preparation for a broader attack. Military leaders then consider how to respond before they have reliable proof that the opposing government ordered the original operation.

An attack is detected before it is known who initiated it

Nuclear command networks link attack-detection devices, centres where political and military leaders assess threats and channels for transmitting orders to nuclear units.

A cyber-attack on that infrastructure can disrupt warning, decision-making or command transmission.

Each of these consequences creates a danger that leaders will mistakenly conclude that an attack has already begun, or that they will soon be unable to respond.

Sisson argues that the decision to launch a cyber-attack on such a network should always be taken by a human.

She would apply the same requirement to operations against infrastructure whose disabling could trigger a serious military crisis, including energy, health, finance and transport.

Such a rule would anchor state responsibility for operations whose consequences could provoke war.

> A military or intelligence system can be taken over by an adversary service, a criminal group, a member of its own organisation or an actor seeking to provoke a conflict between two states

A government that leaves the development and use of such a system to a machine can hardly expect an adversary to accept a later explanation that the attack was not planned.

The possibility of abuse further complicates assessment. A military or intelligence system can be taken over by an adversary service, a criminal group, a member of its own organisation or an actor seeking to provoke a conflict between two states.

As these tools grow in speed and autonomy, there is progressively less time to determine whether an operation was ordered by the government, initiated by an unauthorised user or triggered by a bug in the programme.

Human control has no common definition

Both countries publicly claim that decisions to use military force must remain under human control, but they attach different meanings to the term.

US Department of Defense regulations require autonomous and semi-autonomous systems to be designed so that commanders and operators retain an appropriate level of human judgement when using force.

In a document submitted to the United Nations in 2021, China also supported human responsibility and the ability to control weapons systems, but did not publish sufficiently precise rules to show who makes the final decision, at what stage of an operation and under what conditions.

Washington and Beijing therefore agree on a general principle, although their actual understanding of human control may differ significantly.

The presence of a person in the process does not guarantee that they manage the operation. An officer may formally authorise an attack without sufficient time or data to verify the system's recommendation.

A common definition would have to specify who issues approval, at what stage of the operation, on the basis of what data and with how much time for verification

A commander may have the option to suspend the operation only after the algorithm has already selected the target and begun actions that the adversary perceives as hostile.

A political leader may retain the final say, even though all available options have already been shaped by models whose errors they cannot properly assess.

A common definition would have to specify who issues approval, at what stage of the operation, on the basis of what data and with how much time for verification.

It would also have to include the possibility for a human to stop the system after it has started, as well as clarify responsibility when the behaviour of the programme diverges from its task.

Without these conditions, both governments can claim human control while their armed forces continue to operate under incompatible rules.

The emergency connection must function within the first few minutes

Jiang proposes a dedicated US–China military liaison for incidents involving AI systems.

A government could use that channel to quickly announce that a suspicious operation is the result of a malfunction, tampering or an event still under investigation.

Such a message would not be enough to dispel suspicion immediately, but it would give the other party a reason to delay its response while it investigates what has happened.

Experience with existing channels shows how difficult it would be to implement that proposal.

> For a special relationship to work, its users and their authorisations must be defined in advance

After the downing of a Chinese balloon over the United States in February 2023, the Chinese side did not accept an invitation from the US Secretary of Defense.

Experts on China's decision-making system warn that the lower levels of the military hierarchy have little scope to communicate independently with American counterparts during a sensitive incident.

The response may be delayed while awaiting a decision from the political leadership, even as the military assessment changes from minute to minute.

For a special relationship to work, its users and their authorisations must be defined in advance.

Both sides need to know who is reporting, what information they are allowed to share and what the other army is temporarily suspending while verification is in progress. An additional phone line will not reduce the risk if people on the other end have to wait hours for permission to speak.

A narrow agreement is the only achievable result

A comprehensive US–China agreement on military artificial intelligence currently lacks any political foundation.

Both countries regard the technology as an important source of future military and economic advantage.

Any surveillance that could reveal models, data, command procedures or the capabilities of cyber systems would be unacceptable to both sides.

The scope of the September talks is likely to be limited to a narrow political agreement.

Trump and Xi could reaffirm the 2024 commitment, extend it to cover independent cyber-attacks against nuclear command systems, and instruct officials to agree on the basic meaning of human control

Trump and Xi could reaffirm the 2024 commitment, extend it to cover independent cyber-attacks against nuclear command systems, and instruct officials to agree on the basic meaning of human control.

Establishing a procedure for the immediate notification of a serious malfunction or unauthorised operation would be more achievable than attempting to restrict all military development of AI technology.

Even this outcome is not guaranteed. Bessent and He are engaged in talks fraught with disputes over chips, US export restrictions, open models and Chinese claims that Washington is using security concerns to maintain a technological edge.

The Trump administration is rejecting broader regulation that could slow American companies.

Beijing, for its part, portrays American export restrictions and technological controls as an effort by Washington to preserve its existing advantage under the pretext of national security.

Any possible agreement would regulate crisis management without limiting the development of military algorithms or providing mutual insight into their capabilities.

Its success would be measured by the behaviour of both governments in the first minutes of an incident, when they must determine whether it was a technical failure, an unauthorised action or an ordered attack.

Washington and Beijing can hardly achieve more at the moment, but even such a limited mechanism would reduce the risk that a software problem could one day become a cause of war.