Barriers and Pathways to Human-AI Alignment: A Game-Theoretic Approach

Aran Nayebi
Assistant Professor
Machine Learning Department
Neuroscience Institute (core faculty), Robotics Institute (by courtesy)

Foresight Institute
2025.05.21

ML
MACHINE LEARNING
DEPARTMENT

Carnegie Mellon University
Neuroscience Institute

THE ROBOTICS INSTITUTE

Carnegie Mellon
SCHOOL OF COMPUTER SCIENCE

Operating Principle

If something is *already* inefficient in the theoretically ideal setting of capable agents, then we should avoid it in current practice where we will have malfunctioning or non-cooperative (& non-rational) agents.

Operating Principle

If something is *already* inefficient in the theoretically ideal setting of capable agents, then we should avoid it in current practice where we will have malfunctioning or non-cooperative (& non-rational) agents.

I will show today that we run into several fundamental inefficiencies for AI alignment in general with capable agents.

Motivating Question

Motivating Question

Q: If an agent is “sufficiently capable”, can we also ensure that it is “sufficiently safe”?

Motivating Question

Q: If an agent is "sufficiently capable", can we also ensure that it is "sufficiently safe"?

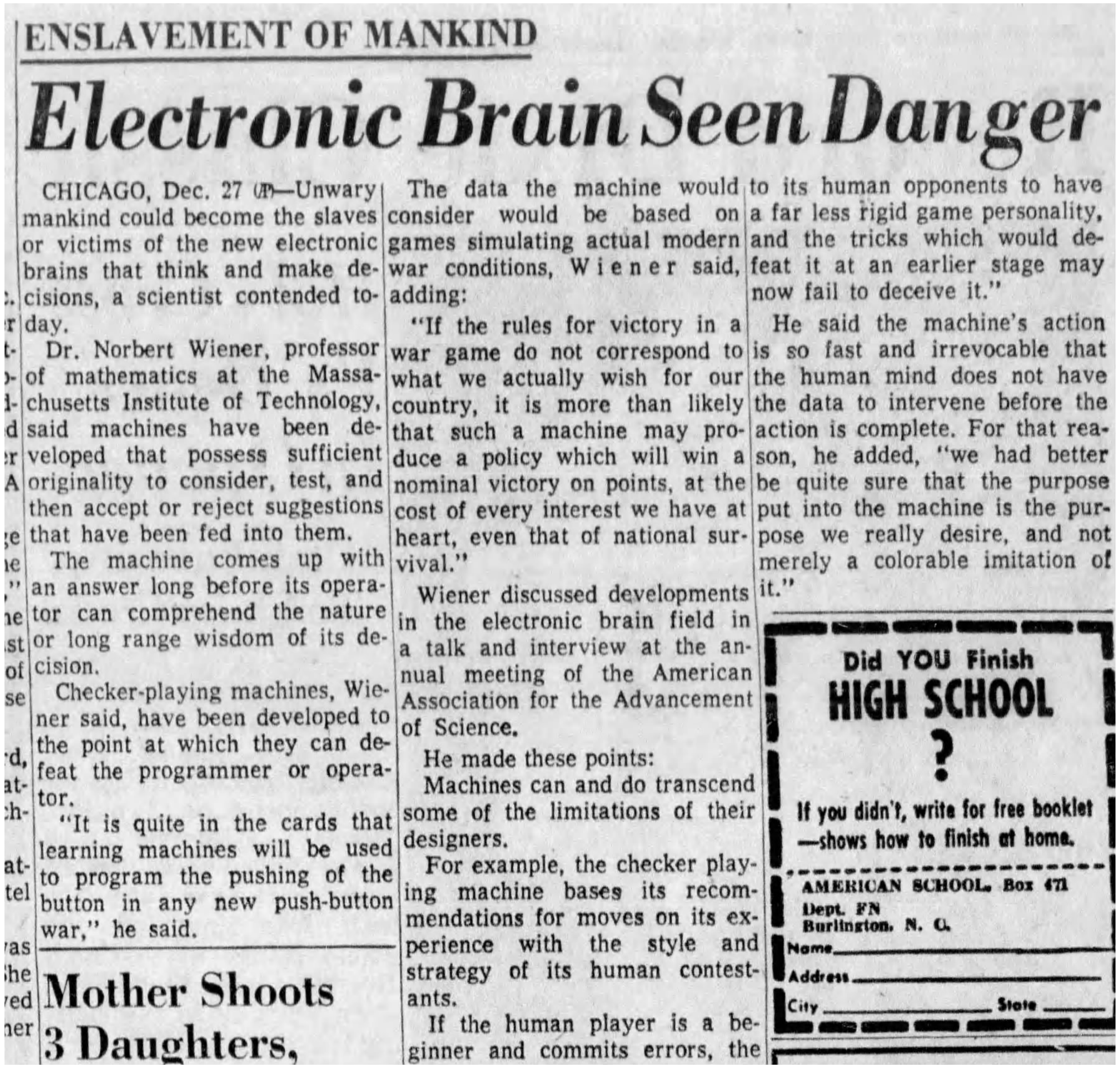
Before Hinton vs. LeCun 2024 there was…

Motivating Question

Q: If an agent is "sufficiently capable", can we also ensure that it is "sufficiently safe"?

Before Hinton vs. LeCun 2024 there was…

Norbert Weiner (MIT) vs. Arthur Samuel (IBM) 1959



ENSLAVEMENT OF MANKIND

Electronic Brain Seen Danger

CHICAGO, Dec. 27 (AP)—Unwary mankind could become the slaves or victims of the new electronic brains that think and make decisions, a scientist contended today.

Dr. Norbert Wiener, professor of mathematics at the Massachusetts Institute of Technology, said machines have been developed that possess sufficient originality to consider, test, and then accept or reject suggestions that have been fed into them.

The machine comes up with an answer long before its operator can comprehend the nature or long range wisdom of its decision.

Checker-playing machines, Wiener said, have been developed to the point at which they can defeat the programmer or operator.

"It is quite in the cards that learning machines will be used to program the pushing of the button in any new push-button war," he said.

The data the machine would consider would be based on games simulating actual modern war conditions, Wiener said, adding:

"If the rules for victory in a war game do not correspond to what we actually wish for our country, it is more than likely that such a machine may produce a policy which will win a nominal victory on points, at the cost of every interest we have at heart, even that of national survival."

Wiener discussed developments in the electronic brain field in a talk and interview at the annual meeting of the American Association for the Advancement of Science.

He made these points:

Machines can and do transcend some of the limitations of their designers.

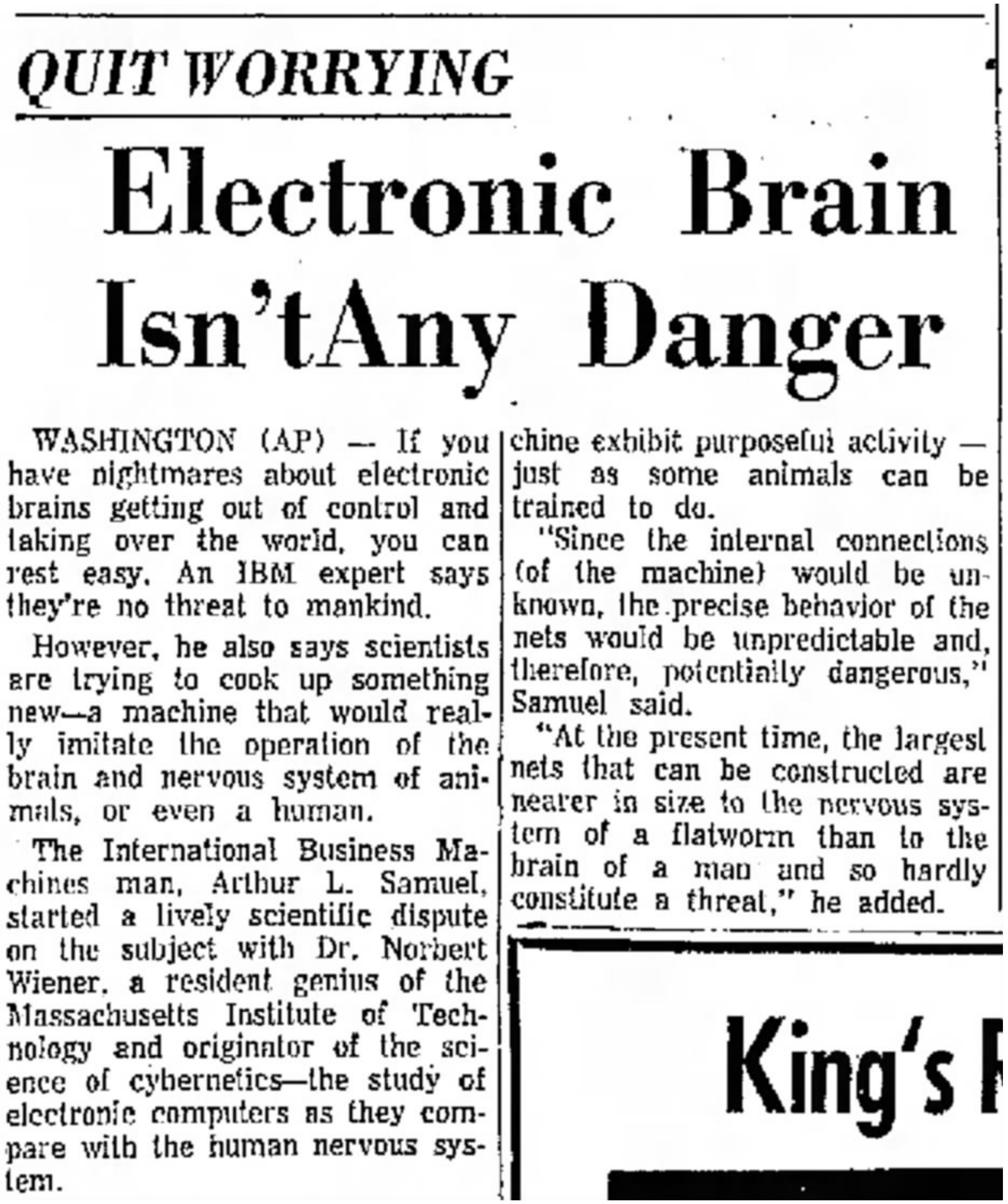
For example, the checker playing machine bases its recommendations for moves on its experience with the style and strategy of its human contestants.

If the human player is a beginner and commits errors, the

to its human opponents to have a far less rigid game personality, and the tricks which would defeat it at an earlier stage may now fail to deceive it."

He said the machine's action is so fast and irrevocable that the human mind does not have the data to intervene before the action is complete. For that reason, he added, "we had better be quite sure that the purpose put into the machine is the purpose we really desire, and not merely a colorable imitation of it."

Mother Shoots 3 Daughters,



QUIT WORRYING

Electronic Brain Isn'tAny Danger

WASHINGTON (AP) — If you have nightmares about electronic brains getting out of control and taking over the world, you can rest easy. An IBM expert says they're no threat to mankind.

However, he also says scientists are trying to cook up something new—a machine that would really imitate the operation of the brain and nervous system of animals, or even a human.

The International Business Machines man, Arthur L. Samuel, started a lively scientific dispute on the subject with Dr. Norbert Wiener, a resident genius of the Massachusetts Institute of Technology and originator of the science of cybernetics—the study of electronic computers as they compare with the human nervous system.

chine exhibit purposeful activity — just as some animals can be trained to do.

"Since the internal connections (of the machine) would be unknown, the precise behavior of the nets would be unpredictable and, therefore, potentially dangerous," Samuel said.

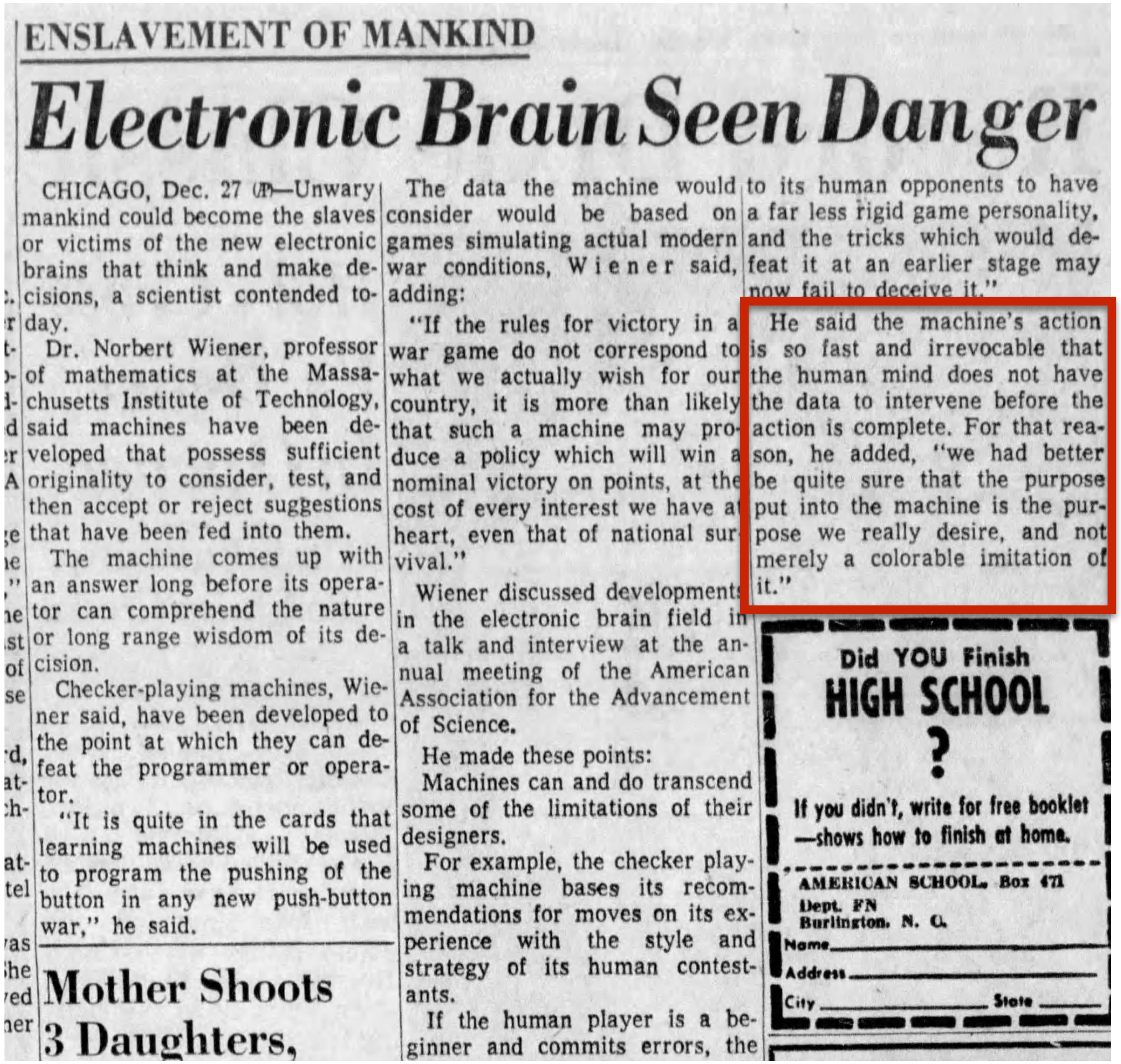
"At the present time, the largest nets that can be constructed are nearer in size to the nervous system of a flatworm than to the brain of a man and so hardly constitute a threat," he added.

Motivating Question

Q: If an agent is "sufficiently capable", can we also ensure that it is "sufficiently safe"?

Before Hinton vs. LeCun 2024 there was…

Norbert Weiner (MIT) vs. Arthur Samuel (IBM) 1959



ENSLAVEMENT OF MANKIND

Electronic Brain Seen Danger

CHICAGO, Dec. 27 (AP)—Unwary mankind could become the slaves or victims of the new electronic brains that think and make decisions, a scientist contended today.

Dr. Norbert Wiener, professor of mathematics at the Massachusetts Institute of Technology, said machines have been developed that possess sufficient originality to consider, test, and then accept or reject suggestions that have been fed into them.

The machine comes up with an answer long before its operator can comprehend the nature or long range wisdom of its decision.

Checker-playing machines, Wiener said, have been developed to the point at which they can defeat the programmer or operator.

"It is quite in the cards that learning machines will be used to program the pushing of the button in any new push-button war," he said.

The data the machine would consider would be based on games simulating actual modern war conditions, Wiener said, adding:

"If the rules for victory in a war game do not correspond to what we actually wish for our country, it is more than likely that such a machine may produce a policy which will win a nominal victory on points, at the cost of every interest we have at heart, even that of national survival."

Wiener discussed developments in the electronic brain field in a talk and interview at the annual meeting of the American Association for the Advancement of Science.

He made these points:

Machines can and do transcend some of the limitations of their designers.

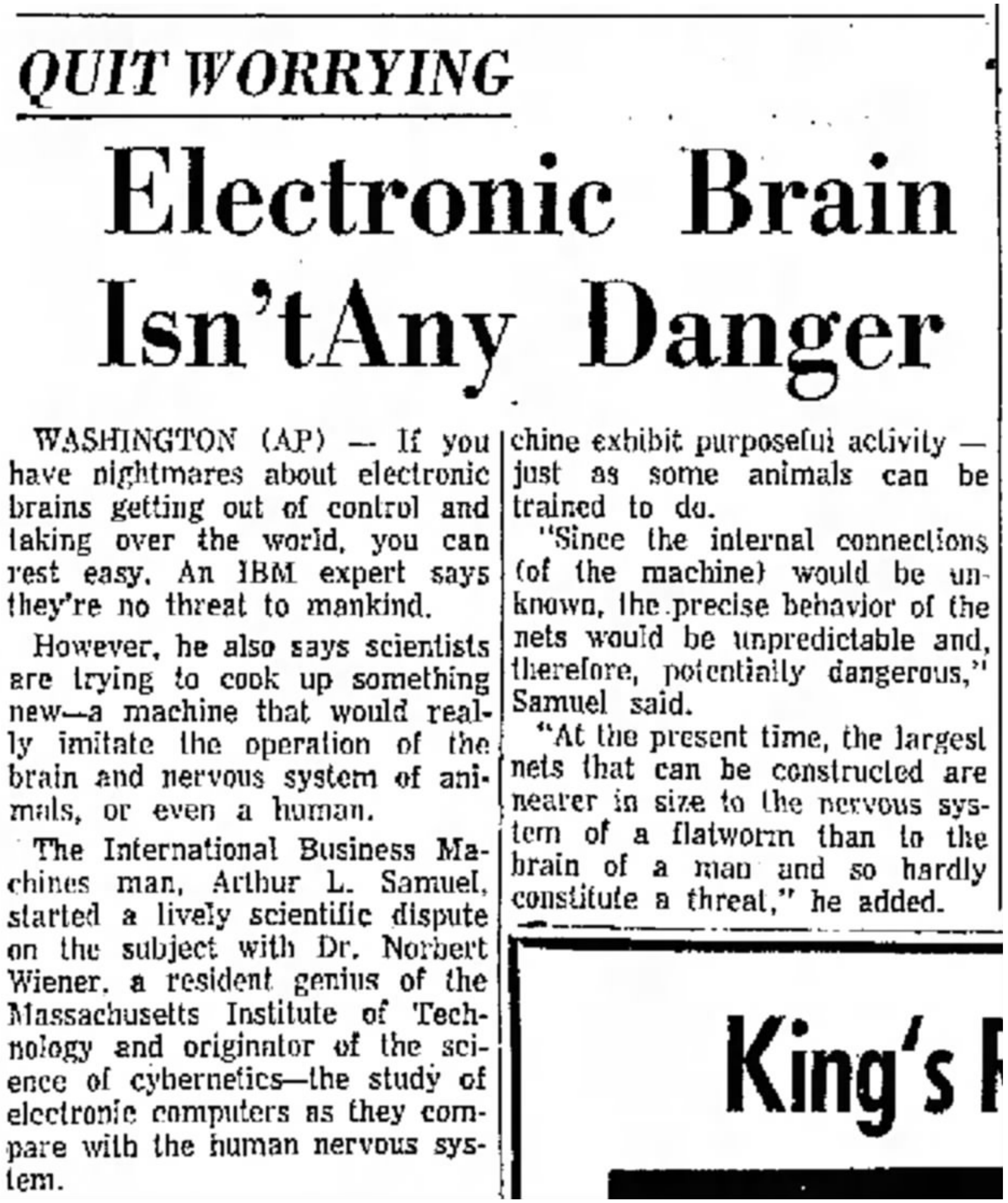
For example, the checker playing machine bases its recommendations for moves on its experience with the style and strategy of its human contestants.

If the human player is a beginner and commits errors, the

to its human opponents to have a far less rigid game personality, and the tricks which would defeat it at an earlier stage may now fail to deceive it."

He said the machine's action is so fast and irrevocable that the human mind does not have the data to intervene before the action is complete. For that reason, he added, "we had better be quite sure that the purpose put into the machine is the purpose we really desire, and not merely a colorable imitation of it."

Mother Shoots 3 Daughters,



QUIT WORRYING

Electronic Brain Isn'tAny Danger

WASHINGTON (AP) — If you have nightmares about electronic brains getting out of control and taking over the world, you can rest easy. An IBM expert says they're no threat to mankind.

However, he also says scientists are trying to cook up something new—a machine that would really imitate the operation of the brain and nervous system of animals, or even a human.

The International Business Machines man, Arthur L. Samuel, started a lively scientific dispute on the subject with Dr. Norbert Wiener, a resident genius of the Massachusetts Institute of Technology and originator of the science of cybernetics—the study of electronic computers as they compare with the human nervous system.

chine exhibit purposeful activity — just as some animals can be trained to do.

"Since the internal connections (of the machine) would be unknown, the precise behavior of the nets would be unpredictable and, therefore, potentially dangerous," Samuel said.

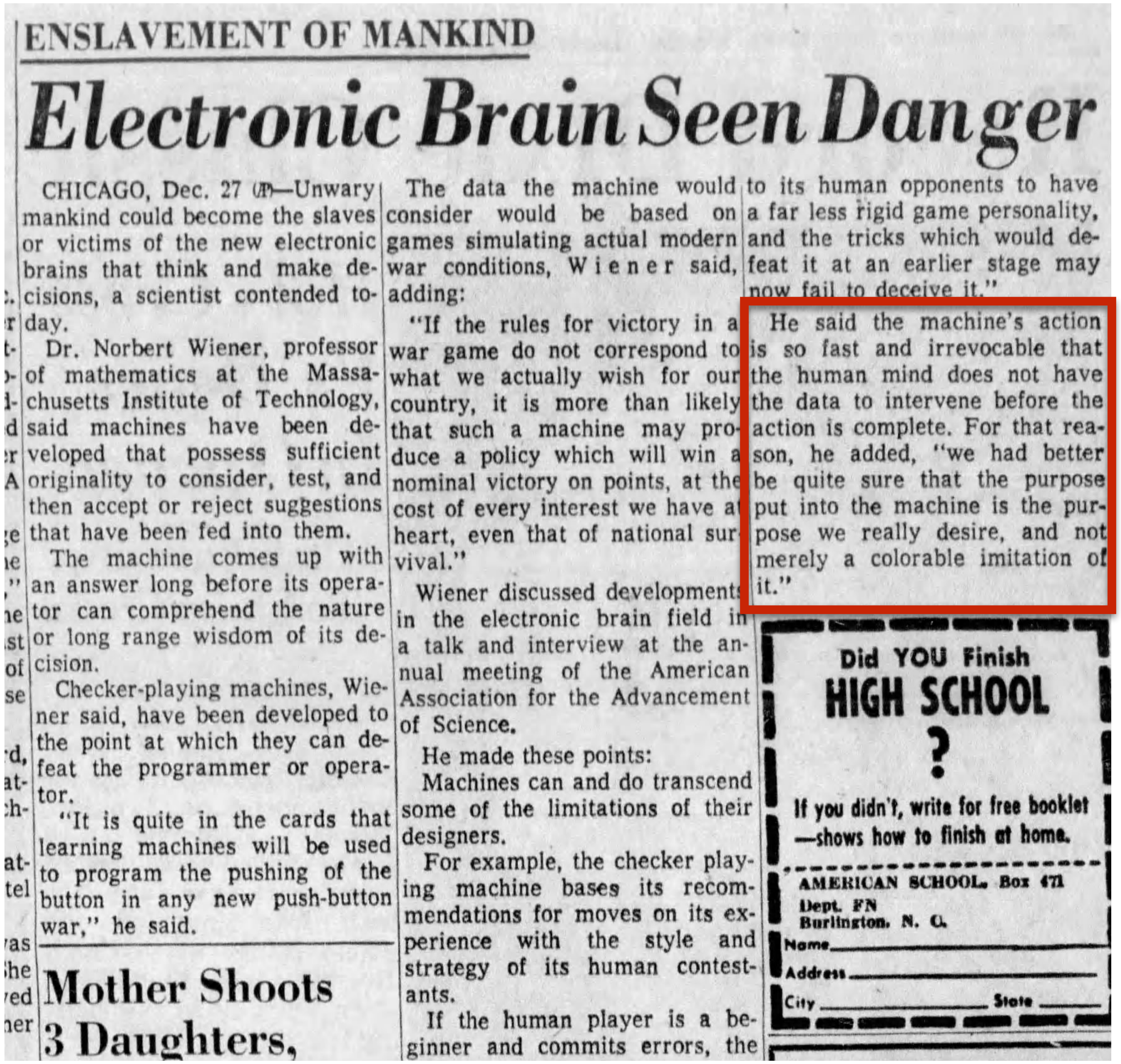
"At the present time, the largest nets that can be constructed are nearer in size to the nervous system of a flatworm than to the brain of a man and so hardly constitute a threat," he added.

Motivating Question

Q: If an agent is "sufficiently capable", can we also ensure that it is "sufficiently safe"?

Before Hinton vs. LeCun 2024 there was…

Norbert Weiner (MIT) vs. Arthur Samuel (IBM) 1959



ENSLAVEMENT OF MANKIND

Electronic Brain Seen Danger

CHICAGO, Dec. 27 (AP)—Unwary mankind could become the slaves or victims of the new electronic brains that think and make decisions, a scientist contended today.

Dr. Norbert Wiener, professor of mathematics at the Massachusetts Institute of Technology, said machines have been developed that possess sufficient originality to consider, test, and then accept or reject suggestions that have been fed into them.

The machine comes up with an answer long before its operator can comprehend the nature or long range wisdom of its decision.

Checker-playing machines, Wiener said, have been developed to the point at which they can defeat the programmer or operator.

"It is quite in the cards that learning machines will be used to program the pushing of the button in any new push-button war," he said.

The data the machine would consider would be based on games simulating actual modern war conditions, Wiener said, adding:

"If the rules for victory in a war game do not correspond to what we actually wish for our country, it is more than likely that such a machine may produce a policy which will win a nominal victory on points, at the cost of every interest we have at heart, even that of national survival."

Wiener discussed developments in the electronic brain field in a talk and interview at the annual meeting of the American Association for the Advancement of Science.

He made these points:

Machines can and do transcend some of the limitations of their designers.

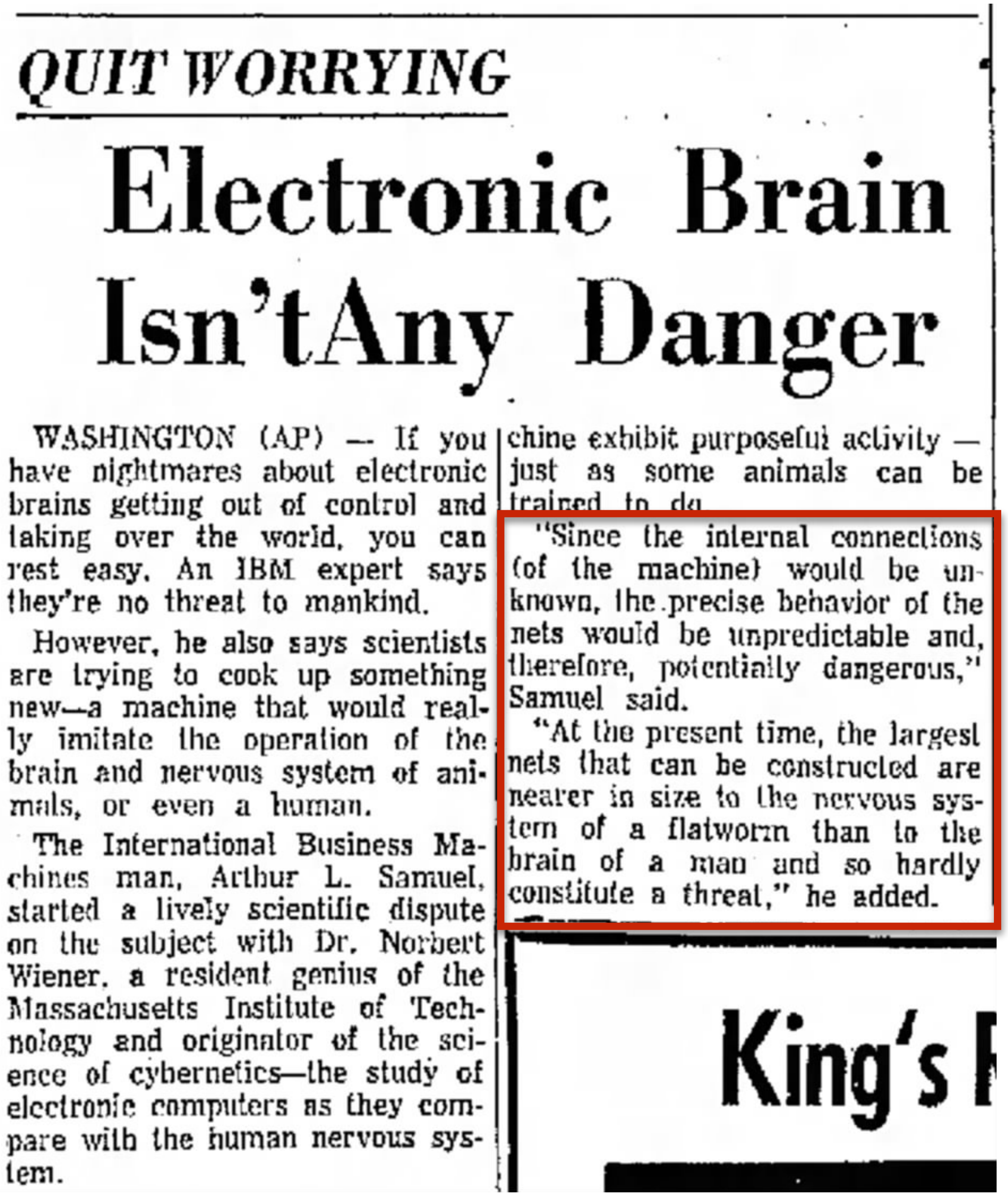
For example, the checker playing machine bases its recommendations for moves on its experience with the style and strategy of its human contestants.

If the human player is a beginner and commits errors, the

to its human opponents to have a far less rigid game personality, and the tricks which would defeat it at an earlier stage may now fail to deceive it."

He said the machine's action is so fast and irrevocable that the human mind does not have the data to intervene before the action is complete. For that reason, he added, "we had better be quite sure that the purpose put into the machine is the purpose we really desire, and not merely a colorable imitation of it."

Mother Shoots 3 Daughters,



QUIT WORRYING

Electronic Brain Isn'tAny Danger

WASHINGTON (AP) — If you have nightmares about electronic brains getting out of control and taking over the world, you can rest easy. An IBM expert says they're no threat to mankind.

However, he also says scientists are trying to cook up something new—a machine that would really imitate the operation of the brain and nervous system of animals, or even a human.

The International Business Machines man, Arthur L. Samuel, started a lively scientific dispute on the subject with Dr. Norbert Wiener, a resident genius of the Massachusetts Institute of Technology and originator of the science of cybernetics—the study of electronic computers as they compare with the human nervous system.

chine exhibit purposeful activity — just as some animals can be trained to do.

"Since the internal connections (of the machine) would be unknown, the precise behavior of the nets would be unpredictable and, therefore, potentially dangerous," Samuel said.

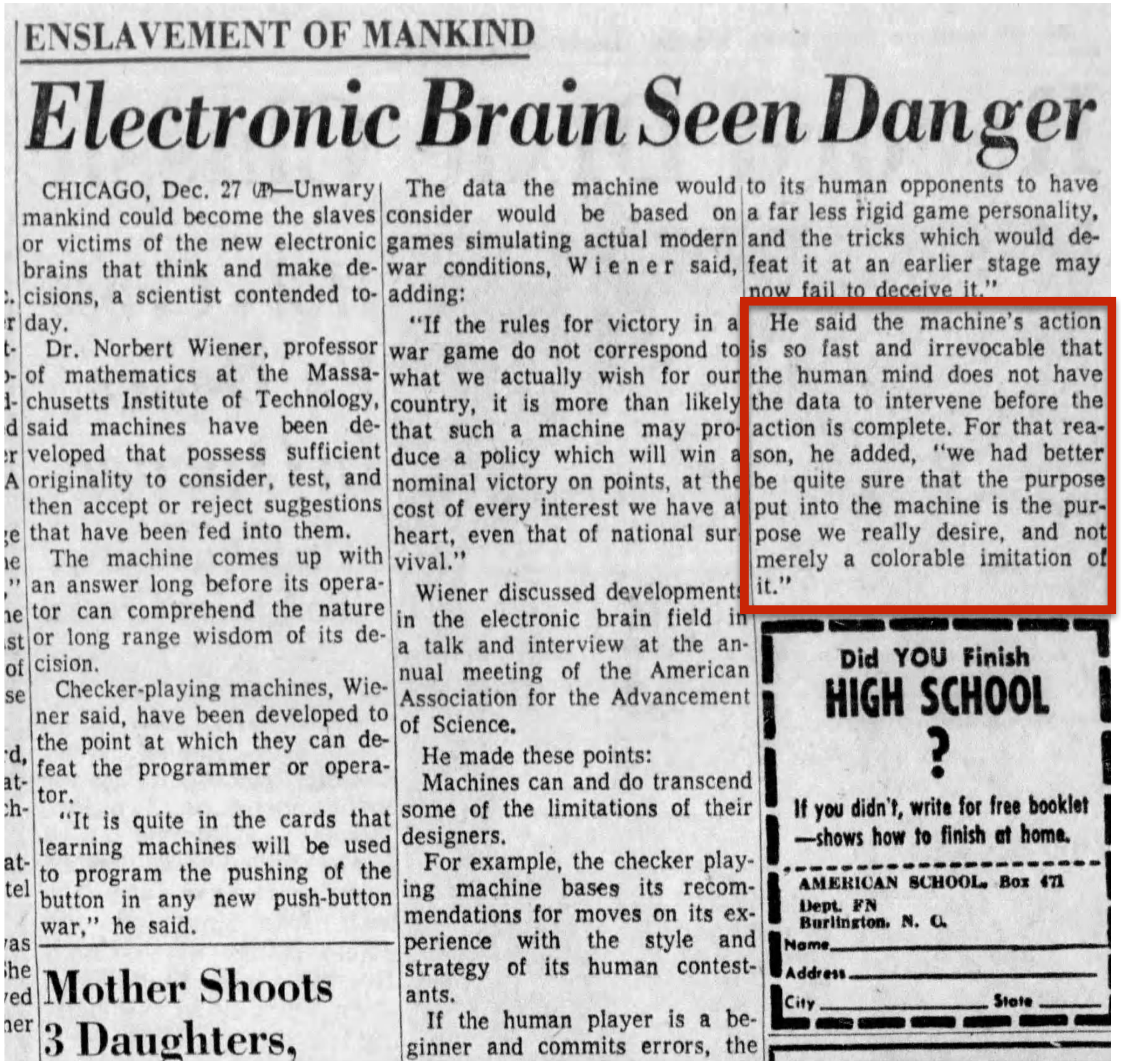
"At the present time, the largest nets that can be constructed are nearer in size to the nervous system of a flatworm than to the brain of a man and so hardly constitute a threat," he added.

Motivating Question

Q: If an agent is "sufficiently capable", can we also ensure that it is "sufficiently safe"?

Before Hinton vs. LeCun 2024 there was…

Norbert Weiner (MIT) vs. Arthur Samuel (IBM) 1959



ENSLAVEMENT OF MANKIND

Electronic Brain Seen Danger

CHICAGO, Dec. 27 (AP)—Unwary mankind could become the slaves or victims of the new electronic brains that think and make decisions, a scientist contended today.

Dr. Norbert Wiener, professor of mathematics at the Massachusetts Institute of Technology, said machines have been developed that possess sufficient originality to consider, test, and then accept or reject suggestions that have been fed into them.

The machine comes up with an answer long before its operator can comprehend the nature or long range wisdom of its decision.

Checker-playing machines, Wiener said, have been developed to the point at which they can defeat the programmer or operator.

"It is quite in the cards that learning machines will be used to program the pushing of the button in any new push-button war," he said.

The data the machine would consider would be based on games simulating actual modern war conditions, Wiener said, adding:

"If the rules for victory in a war game do not correspond to what we actually wish for our country, it is more than likely that such a machine may produce a policy which will win a nominal victory on points, at the cost of every interest we have at heart, even that of national survival."

Wiener discussed developments in the electronic brain field in a talk and interview at the annual meeting of the American Association for the Advancement of Science.

He made these points:

Machines can and do transcend some of the limitations of their designers.

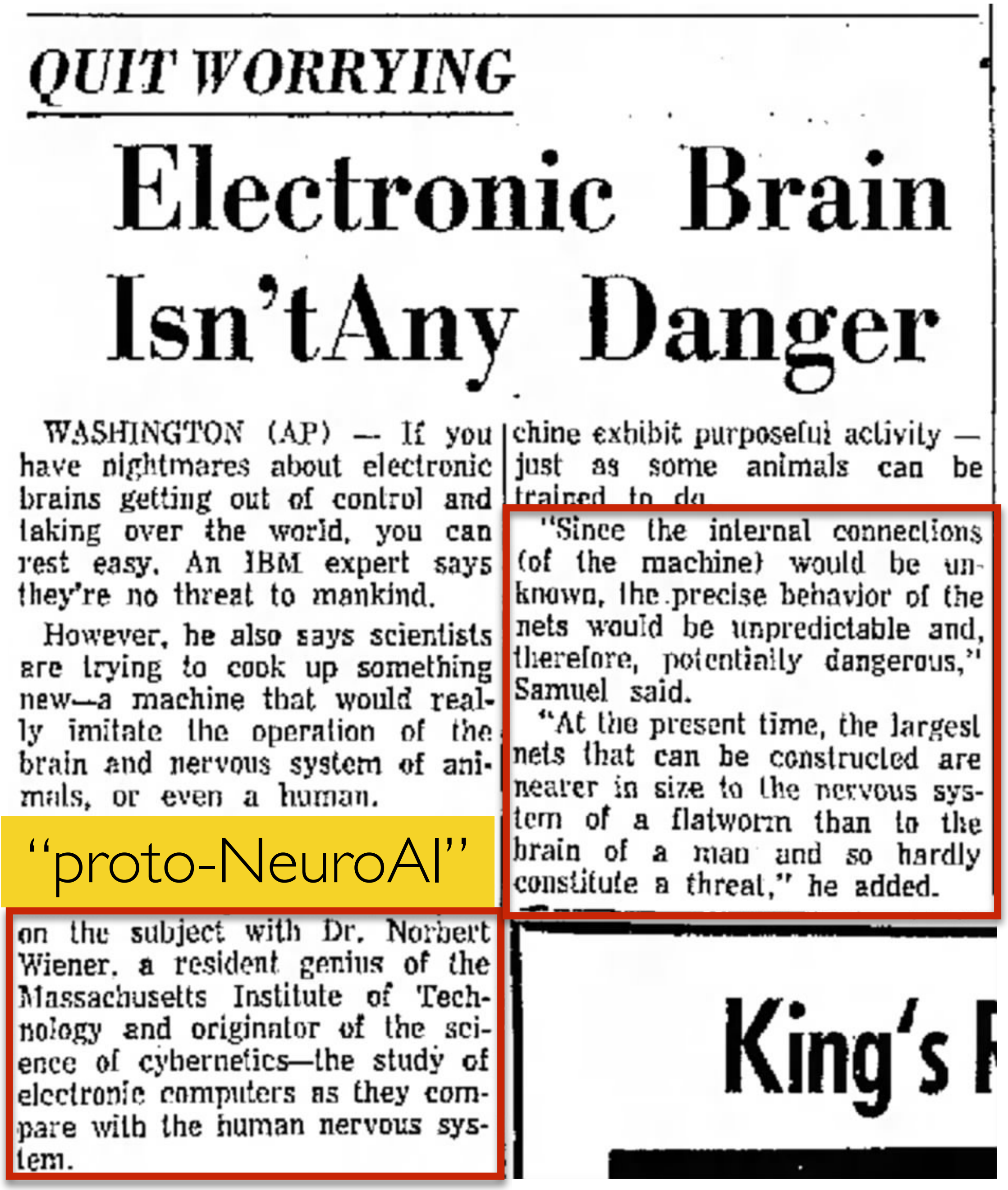
For example, the checker playing machine bases its recommendations for moves on its experience with the style and strategy of its human contestants.

If the human player is a beginner and commits errors, the

to its human opponents to have a far less rigid game personality, and the tricks which would defeat it at an earlier stage may now fail to deceive it."

He said the machine's action is so fast and irrevocable that the human mind does not have the data to intervene before the action is complete. For that reason, he added, "we had better be quite sure that the purpose put into the machine is the purpose we really desire, and not merely a colorable imitation of it."

Mother Shoots 3 Daughters,



QUIT WORRYING

Electronic Brain Isn'tAny Danger

WASHINGTON (AP) — If you have nightmares about electronic brains getting out of control and taking over the world, you can rest easy. An IBM expert says they're no threat to mankind.

However, he also says scientists are trying to cook up something new—a machine that would really imitate the operation of the brain and nervous system of animals, or even a human.

on the subject with Dr. Norbert Wiener, a resident genius of the Massachusetts Institute of Technology and originator of the science of cybernetics—the study of electronic computers as they compare with the human nervous system.

chine exhibit purposeful activity — just as some animals can be trained to do.

"Since the internal connections (of the machine) would be unknown, the precise behavior of the nets would be unpredictable and, therefore, potentially dangerous," Samuel said.

"At the present time, the largest nets that can be constructed are nearer in size to the nervous system of a flatworm than to the brain of a man and so hardly constitute a threat," he added.

King's

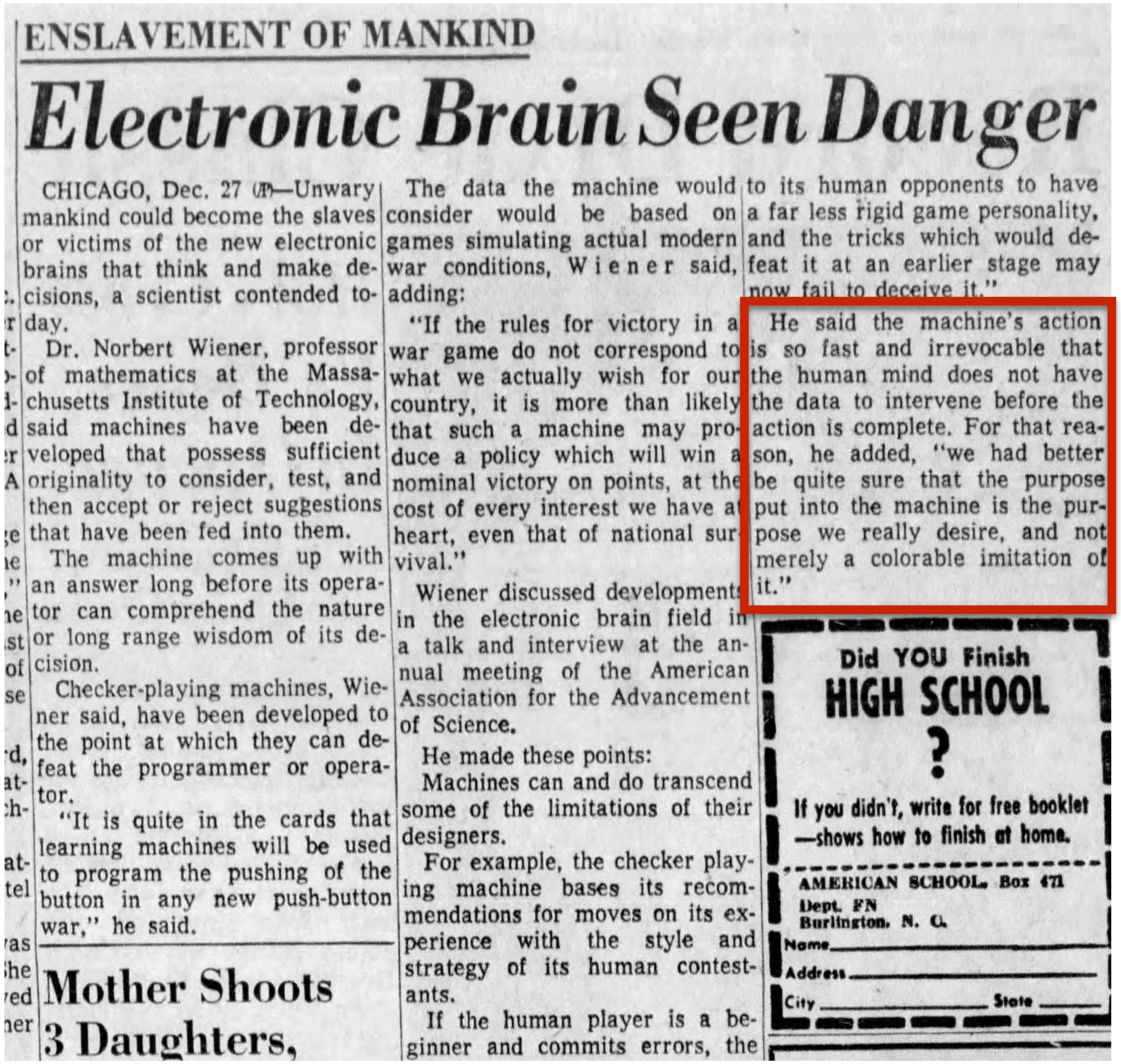
"proto-NeuroAI"

Motivating Question

Q: If an agent is “sufficiently capable”, can we also ensure that it is “sufficiently safe”? What do these words mean?

Before Hinton vs. LeCun 2024 there was…

Norbert Weiner (MIT) vs. Arthur Samuel (IBM) 1959



ENSLAVEMENT OF MANKIND

Electronic Brain Seen Danger

CHICAGO, Dec. 27 (AP)—Unwary mankind could become the slaves or victims of the new electronic brains that think and make decisions, a scientist contended today.

Dr. Norbert Wiener, professor of mathematics at the Massachusetts Institute of Technology, said machines have been developed that possess sufficient originality to consider, test, and then accept or reject suggestions that have been fed into them.

The machine comes up with an answer long before its operator can comprehend the nature or long range wisdom of its decision.

Checker-playing machines, Wiener said, have been developed to the point at which they can defeat the programmer or operator.

“It is quite in the cards that learning machines will be used to program the pushing of the button in any new push-button war,” he said.

The data the machine would consider would be based on games simulating actual modern war conditions, Wiener said, adding:

“If the rules for victory in a war game do not correspond to what we actually wish for our country, it is more than likely that such a machine may produce a policy which will win a nominal victory on points, at the cost of every interest we have at heart, even that of national survival.”

Wiener discussed developments in the electronic brain field in a talk and interview at the annual meeting of the American Association for the Advancement of Science.

He made these points:

Machines can and do transcend some of the limitations of their designers.

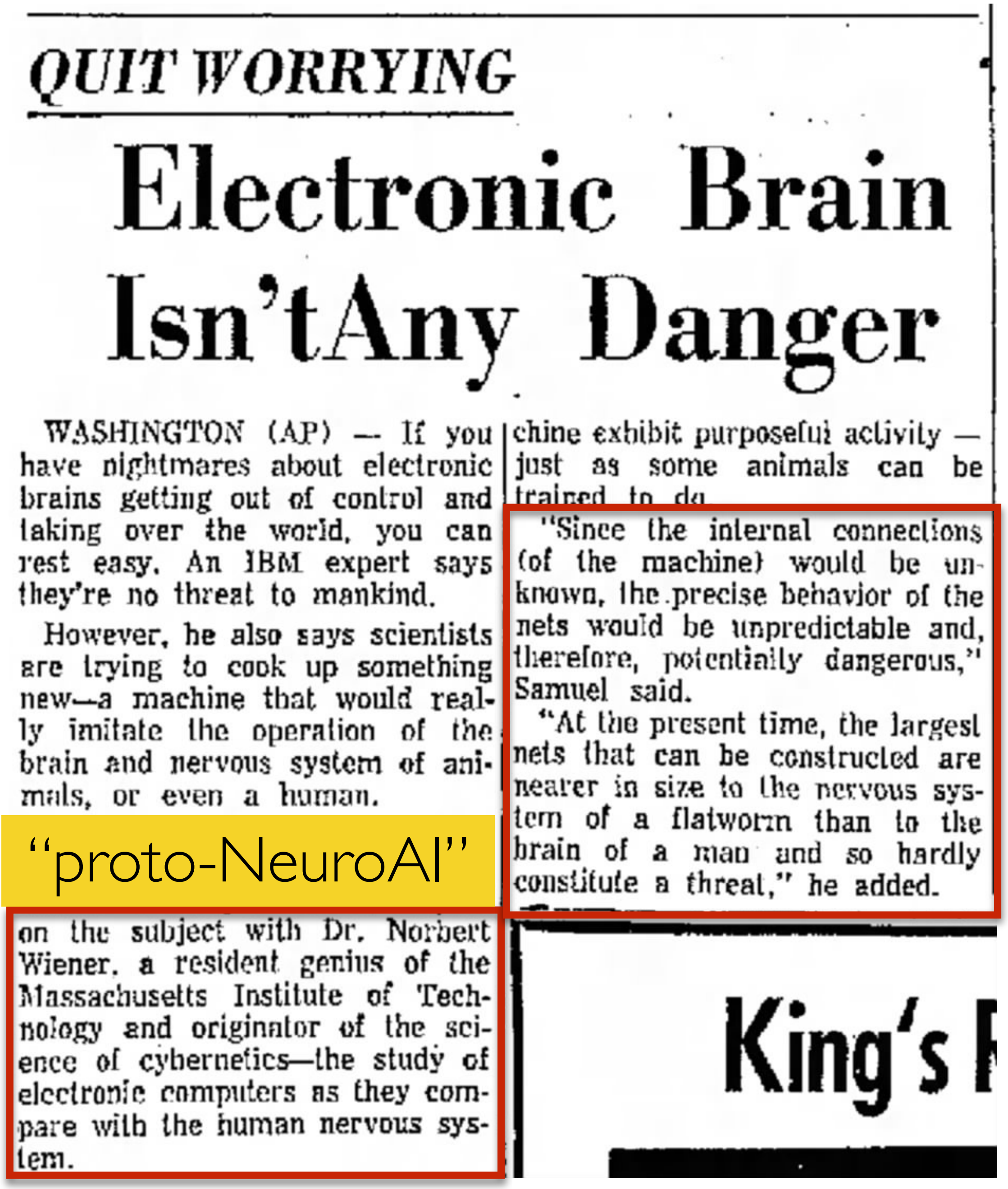
For example, the checker playing machine bases its recommendations for moves on its experience with the style and strategy of its human contestants.

If the human player is a beginner and commits errors, the

to its human opponents to have a far less rigid game personality, and the tricks which would defeat it at an earlier stage may now fail to deceive it.”

He said the machine’s action is so fast and irrevocable that the human mind does not have the data to intervene before the action is complete. For that reason, he added, “we had better be quite sure that the purpose put into the machine is the purpose we really desire, and not merely a colorable imitation of it.”

Mother Shoots 3 Daughters,



QUIT WORRYING

Electronic Brain Isn'tAny Danger

WASHINGTON (AP) — If you have nightmares about electronic brains getting out of control and taking over the world, you can rest easy. An IBM expert says they're no threat to mankind.

However, he also says scientists are trying to cook up something new—a machine that would really imitate the operation of the brain and nervous system of animals, or even a human.

“proto-NeuroAI”

on the subject with Dr. Norbert Wiener, a resident genius of the Massachusetts Institute of Technology and originator of the science of cybernetics—the study of electronic computers as they compare with the human nervous system.

chine exhibit purposeful activity — just as some animals can be trained to do.

“Since the internal connections (of the machine) would be unknown, the precise behavior of the nets would be unpredictable and, therefore, potentially dangerous,” Samuel said.

“At the present time, the largest nets that can be constructed are nearer in size to the nervous system of a flatworm than to the brain of a man and so hardly constitute a threat,” he added.

Lots of Discussion

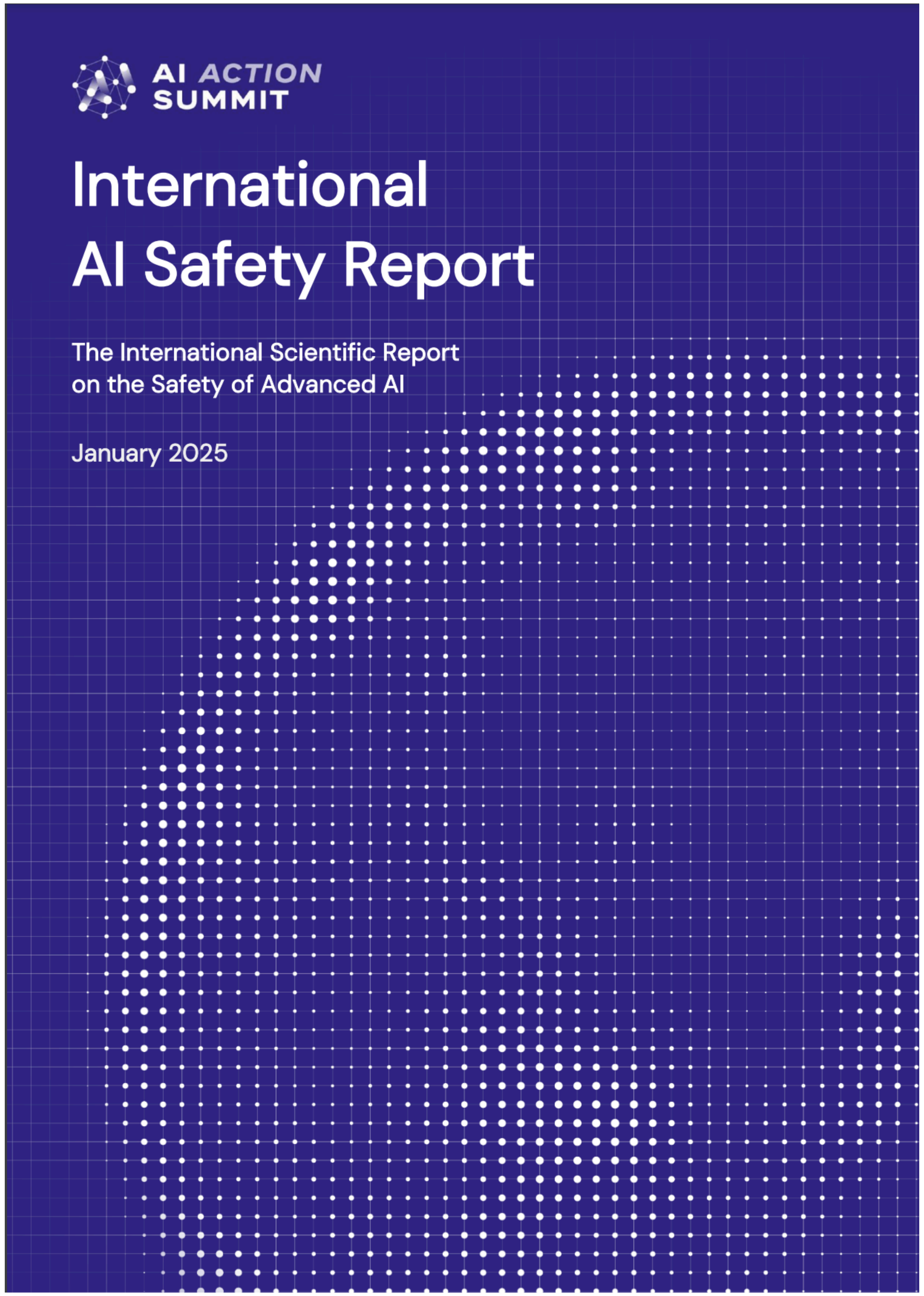
Q: If an agent is "sufficiently capable", can we also ensure that it is "sufficiently safe"?

What do these words mean?

Lots of Discussion

Q: If an agent is “sufficiently capable”, can we also ensure that it is “sufficiently safe”?

What do these words mean?



AI ACTION SUMMIT

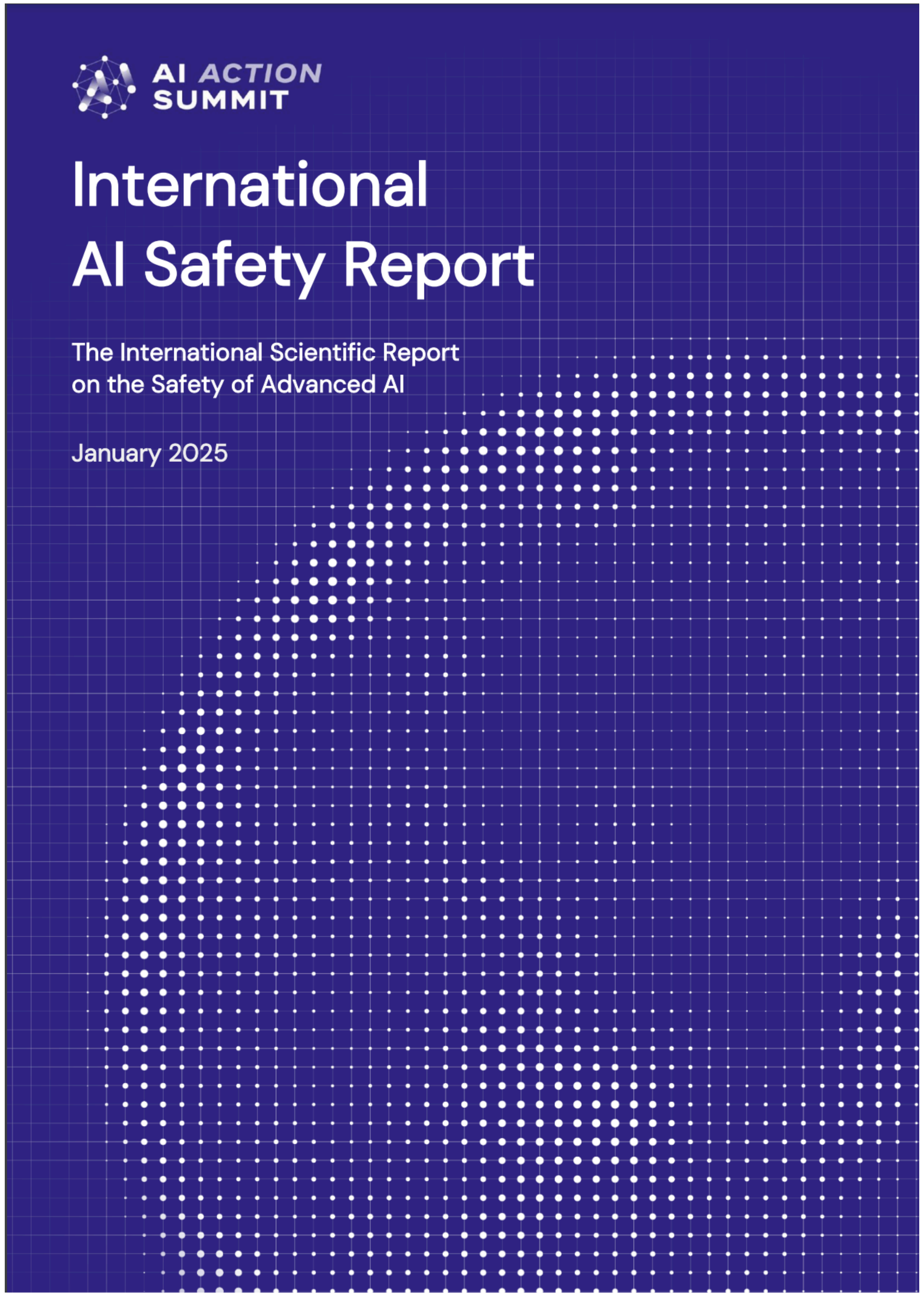
International AI Safety Report

The International Scientific Report on the Safety of Advanced AI

January 2025

Lots of Discussion

Q: If an agent is "sufficiently capable", can we also ensure that it is "sufficiently safe"? What do these words mean?



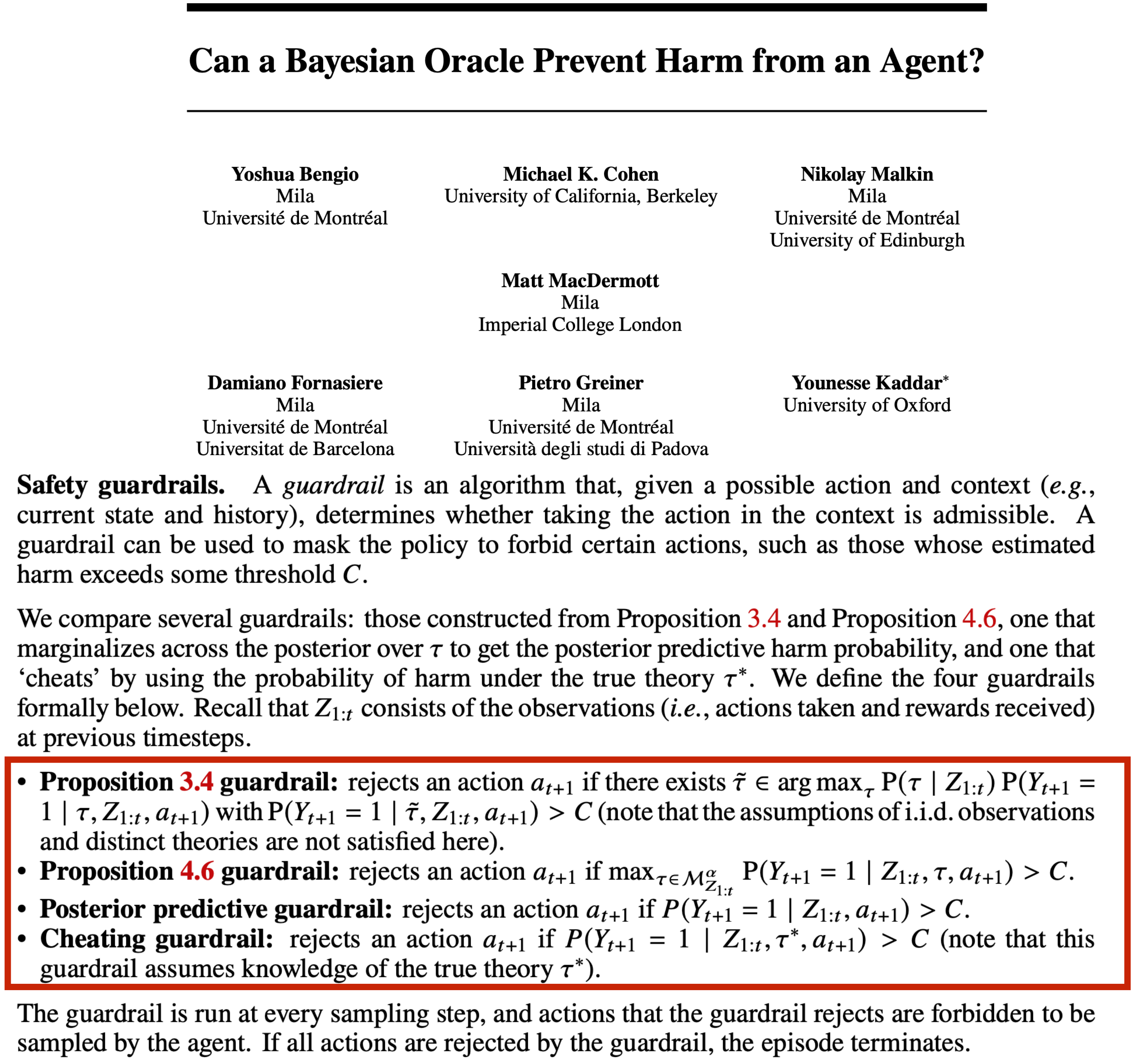
AI ACTION SUMMIT

International AI Safety Report

The International Scientific Report on the Safety of Advanced AI

January 2025

4 different definitions of a guardrail!



Can a Bayesian Oracle Prevent Harm from an Agent?

Yoshua Bengio
Mila
Université de Montréal

Michael K. Cohen
University of California, Berkeley

Nikolay Malkin
Mila
Université de Montréal
University of Edinburgh

Matt MacDermott
Mila
Imperial College London

Damiano Fornasiere
Mila
Université de Montréal
Universitat de Barcelona

Pietro Greiner
Mila
Université de Montréal
Università degli studi di Padova

Younesse Kaddar*
University of Oxford

Safety guardrails. A *guardrail* is an algorithm that, given a possible action and context (*e.g.*, current state and history), determines whether taking the action in the context is admissible. A guardrail can be used to mask the policy to forbid certain actions, such as those whose estimated harm exceeds some threshold C.

We compare several guardrails: those constructed from Proposition 3.4 and Proposition 4.6, one that marginalizes across the posterior over τ to get the posterior predictive harm probability, and one that 'cheats' by using the probability of harm under the true theory τ^*. We define the four guardrails formally below. Recall that $Z_{1:t}$ consists of the observations (*i.e.*, actions taken and rewards received) at previous timesteps.

- **Proposition 3.4 guardrail:** rejects an action a_{t+1} if there exists $\tilde{\tau} \in \arg\max_\tau \mathrm{P}(\tau \mid Z_{1:t})\,\mathrm{P}(Y_{t+1} = 1 \mid \tau, Z_{1:t}, a_{t+1})$ with $\mathrm{P}(Y_{t+1} = 1 \mid \tilde{\tau}, Z_{1:t}, a_{t+1}) > C$ (note that the assumptions of i.i.d. observations and distinct theories are not satisfied here).
- **Proposition 4.6 guardrail:** rejects an action a_{t+1} if $\max_{\tau \in \mathcal{M}^{\alpha}_{Z_{1:t}}} \mathrm{P}(Y_{t+1} = 1 \mid Z_{1:t}, \tau, a_{t+1}) > C$.
- **Posterior predictive guardrail:** rejects an action a_{t+1} if $P(Y_{t+1} = 1 \mid Z_{1:t}, a_{t+1}) > C$.
- **Cheating guardrail:** rejects an action a_{t+1} if $P(Y_{t+1} = 1 \mid Z_{1:t}, \tau^*, a_{t+1}) > C$ (note that this guardrail assumes knowledge of the true theory τ^*).

The guardrail is run at every sampling step, and actions that the guardrail rejects are forbidden to be sampled by the agent. If all actions are rejected by the guardrail, the episode terminates.

Shifting to Alignment: A *Relative* Perspective

Q: If an agent is "sufficiently capable", can we also ensure that it is "sufficiently safe"?

What do these words mean?

Shifting to Alignment: A *Relative* Perspective

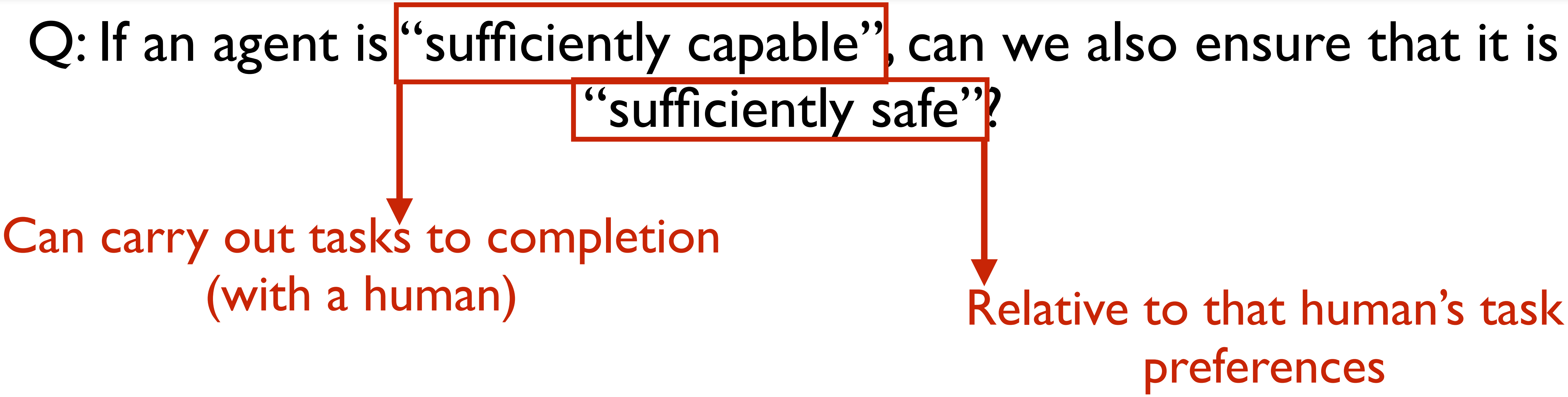
Q: If an agent is "sufficiently capable", can we also ensure that it is "sufficiently safe"?

Shifting to Alignment: A *Relative* Perspective

Q: If an agent is "sufficiently capable", can we also ensure that it is "sufficiently safe"?

Can carry out tasks to completion (with a human)

Shifting to Alignment: A *Relative* Perspective



Q: If an agent is "sufficiently capable", can we also ensure that it is "sufficiently safe"?

Can carry out tasks to completion (with a human)

Relative to that human's task preferences

Shifting to Alignment: A *Relative* Perspective

Q: If an agent is "sufficiently capable", can we also ensure that it is "sufficiently safe"?

Can carry out tasks to completion (with a human)

Relative to that human's task preferences

Q': If an agent can carry out tasks to completion with a human, can we also ensure that it is agrees with the human's preferences for that task?

Shifting to Alignment: A *Relative* Perspective

Q: If an agent is "sufficiently capable", can we also ensure that it is "sufficiently safe"?

Can carry out tasks to completion (with a human)

Relative to that human's task preferences

Q': If an agent can carry out tasks to completion with a human, can we also ensure that it is agrees with the human's preferences for that task?

Note: Current LLM agents don't always satisfy this assumption leading immediately to misalignment, so this would be a quite "generally-capable" agent that's yet to be built.

Shifting to Alignment: A *Relative* Perspective

Q: If an agent is "sufficiently capable", can we also ensure that it is "sufficiently safe"?

Can carry out tasks to completion (with a human)

Relative to that human's task preferences

Q': If an agent can carry out tasks to completion with a human, can we also ensure that it is agrees with the human's preferences for that task?

Note: Current LLM agents don't always satisfy this assumption leading immediately to misalignment, so this would be a quite "generally-capable" agent that's yet to be built.

(Therefore needs a theoretical treatment, since we can't simply "run" these agents forward — we don't have them yet!)

Shifting to Alignment: A *Relative* Perspective

Q: If an agent is "sufficiently capable", can we also ensure that it is "sufficiently safe"?

Can carry out tasks to completion (with a human)

Relative to that human's task preferences

Q': If an agent can carry out tasks to completion with a human, can we also ensure that it is agrees with the human's preferences for that task?

Note: Current LLM agents don't always satisfy this assumption leading immediately to misalignment, so this would be a quite "generally-capable" agent that's yet to be built.

But requires 2 core ingredients:

(1) **Coordination**

(2) **Partial Information**

(Therefore needs a theoretical treatment, since we can't simply "run" these agents forward — we don't have them yet!)

Shifting to Alignment: A *Relative* Perspective

Q: If an agent is "sufficiently capable", can we also ensure that it is "sufficiently safe"?

Can carry out tasks to completion (with a human)

Relative to that human's task preferences

Q': If an agent can carry out tasks to completion with a human, can we also ensure that it is agrees with the human's preferences for that task?

Note: Current LLM agents don't always satisfy this assumption leading immediately to misalignment, so this would be a quite "generally-capable" agent that's yet to be built.

(Therefore needs a theoretical treatment, since we can't simply "run" these agents forward — we don't have them yet!)

But requires 2 core ingredients:

(1) **Coordination**

(2) **Partial Information**

Game Theory!

Shifting to Alignment: A *Relative* Perspective

Q: If an agent is "sufficiently capable", can we also ensure that it is "sufficiently safe"?

Can carry out tasks to completion (with a human)

Relative to that human's task preferences

Q': If an agent can carry out tasks to completion with a human, can we also ensure that it is agrees with the human's preferences for that task?

Note: Current LLM agents don't always satisfy this assumption leading immediately to misalignment, so this would be a quite "generally-capable" agent that's yet to be built.

But requires 2 core ingredients:

(1) **Coordination**

(2) **Partial Information**

Game Theory!

(a bad model of human behavior, but a great model of ideally rational agents)

(Therefore needs a theoretical treatment, since we can't simply "run" these agents forward — we don't have them yet!)

Shifting to Alignment: A *Relative* Perspective

Q: If an agent is "sufficiently capable", can we also ensure that it is "sufficiently safe"?

Can carry out tasks to completion (with a human)

Relative to that human's task preferences

Q': If an agent can carry out tasks to completion with a human, can we also ensure that it is agrees with the human's preferences for that task?

We basically need to formalize this

Note: Current LLM agents don't always satisfy this assumption leading immediately to misalignment, so this would be a quite "generally-capable" agent that's yet to be built.

(Therefore needs a theoretical treatment, since we can't simply "run" these agents forward — we don't have them yet!)

But requires 2 core ingredients:

(1) **Coordination**

(2) **Partial Information**

Game Theory!

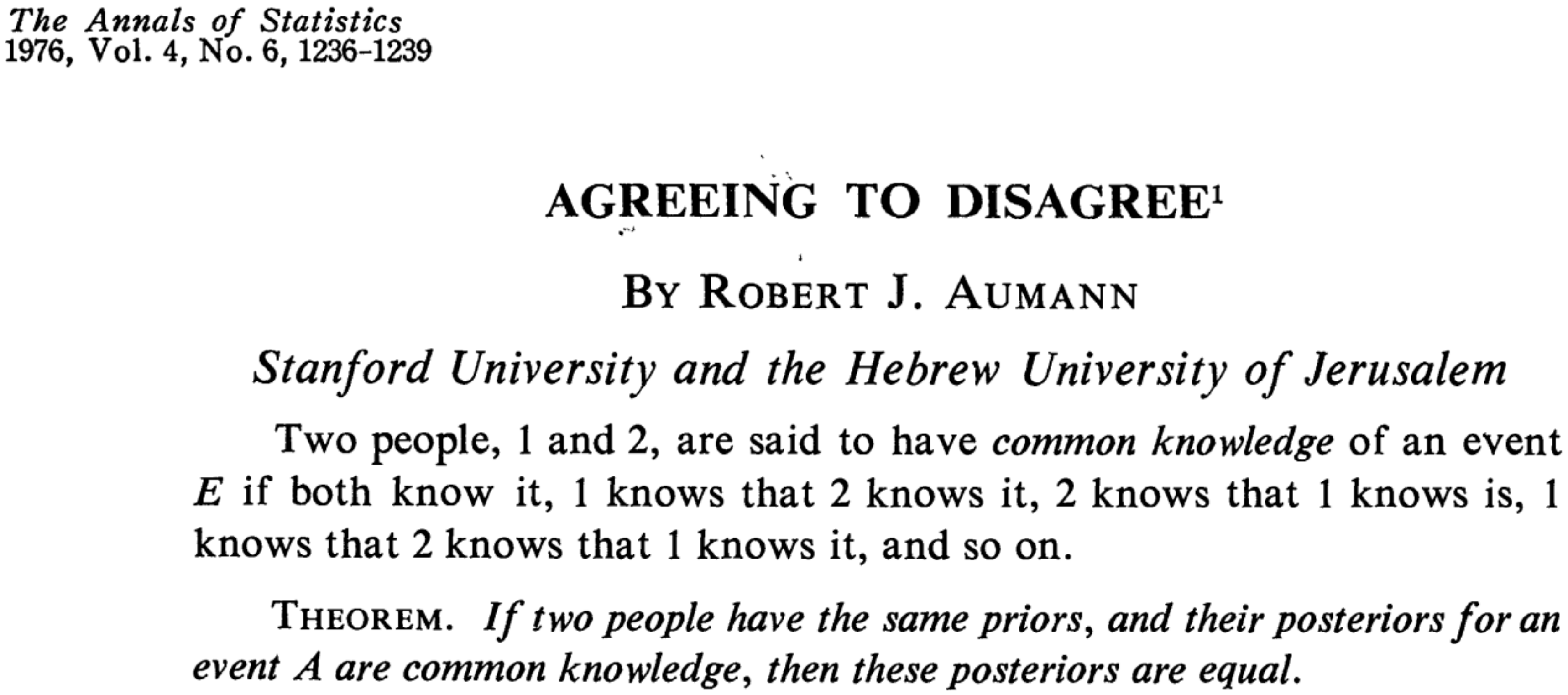
(a bad model of human behavior, but a great model of ideally rational agents)

Aumann's Agreement Theorem

Alice

Bob

Aumann's Agreement Theorem



The Annals of Statistics
1976, Vol. 4, No. 6, 1236–1239

AGREEING TO DISAGREE[1]

BY ROBERT J. AUMANN

Stanford University and the Hebrew University of Jerusalem

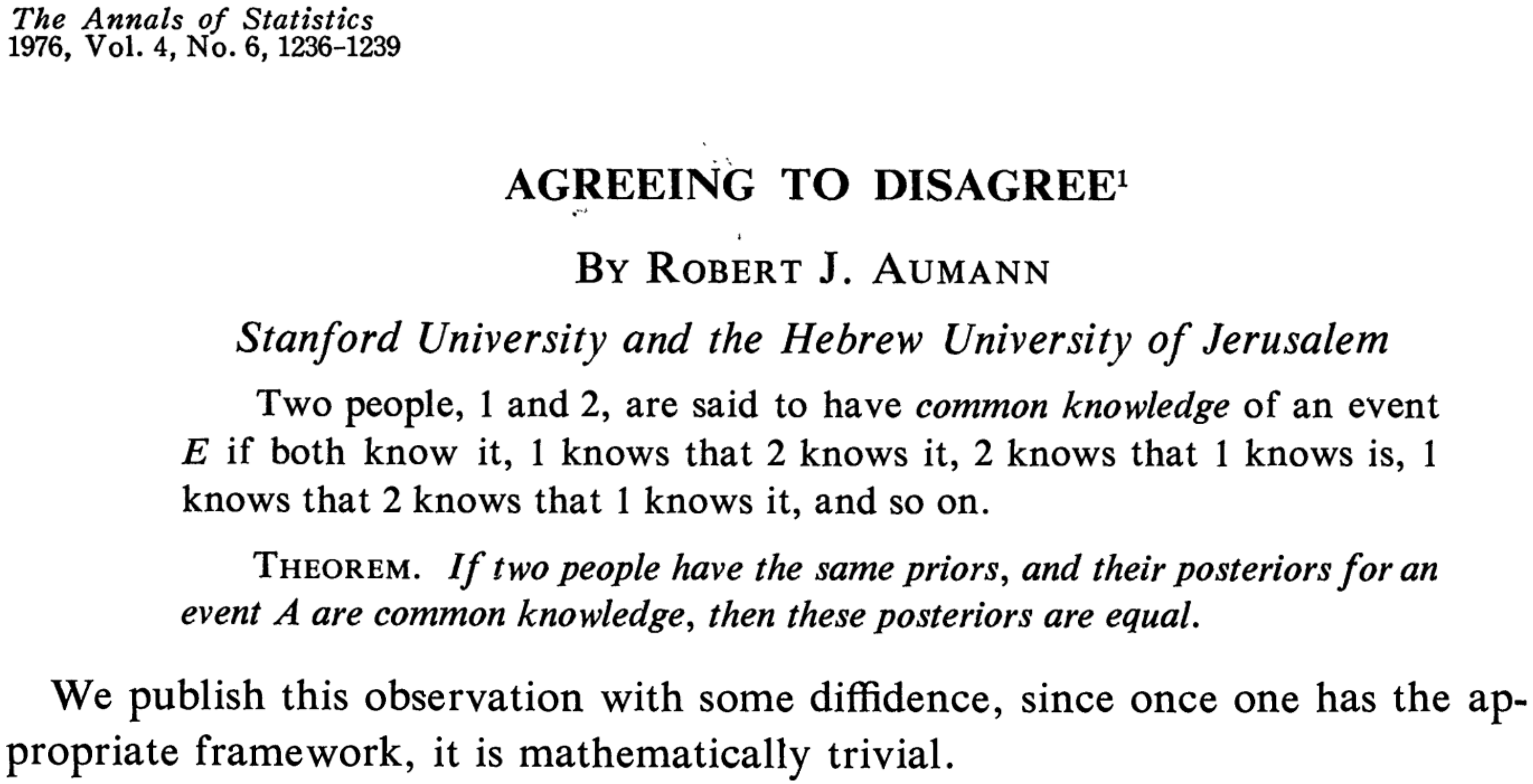
Two people, 1 and 2, are said to have *common knowledge* of an event E if both know it, 1 knows that 2 knows it, 2 knows that 1 knows is, 1 knows that 2 knows that 1 knows it, and so on.

THEOREM. *If two people have the same priors, and their posteriors for an event A are common knowledge, then these posteriors are equal.*

Alice

Bob

Aumann's Agreement Theorem



The Annals of Statistics
1976, Vol. 4, No. 6, 1236–1239

AGREEING TO DISAGREE[1]

BY ROBERT J. AUMANN

Stanford University and the Hebrew University of Jerusalem

Two people, 1 and 2, are said to have *common knowledge* of an event *E* if both know it, 1 knows that 2 knows it, 2 knows that 1 knows is, 1 knows that 2 knows that 1 knows it, and so on.

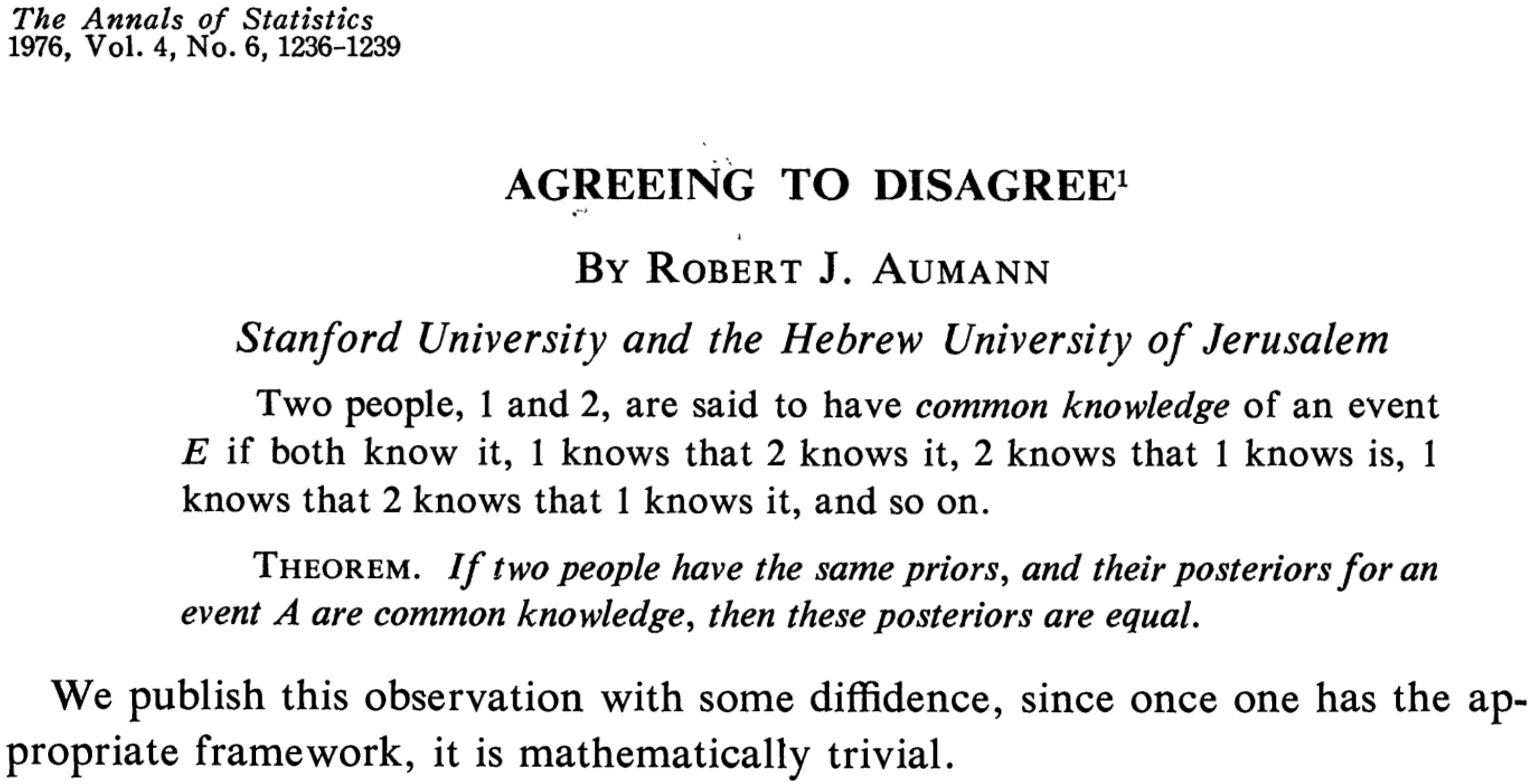
THEOREM. *If two people have the same priors, and their posteriors for an event A are common knowledge, then these posteriors are equal.*

We publish this observation with some diffidence, since once one has the appropriate framework, it is mathematically trivial.

Alice

Bob

Aumann's Agreement Theorem



The Annals of Statistics
1976, Vol. 4, No. 6, 1236–1239

AGREEING TO DISAGREE[1]

By Robert J. Aumann

Stanford University and the Hebrew University of Jerusalem

Two people, 1 and 2, are said to have *common knowledge* of an event E if both know it, 1 knows that 2 knows it, 2 knows that 1 knows is, 1 knows that 2 knows that 1 knows it, and so on.

Theorem. *If two people have the same priors, and their posteriors for an event A are common knowledge, then these posteriors are equal.*

We publish this observation with some diffidence, since once one has the appropriate framework, it is mathematically trivial.

1. Requires a quite strong common prior btwn agents

2. *Humans* are unrealistically modeled as Bayesians

3. Does not study how long the coordination will take (in terms of number of messages)

Alice

Bob

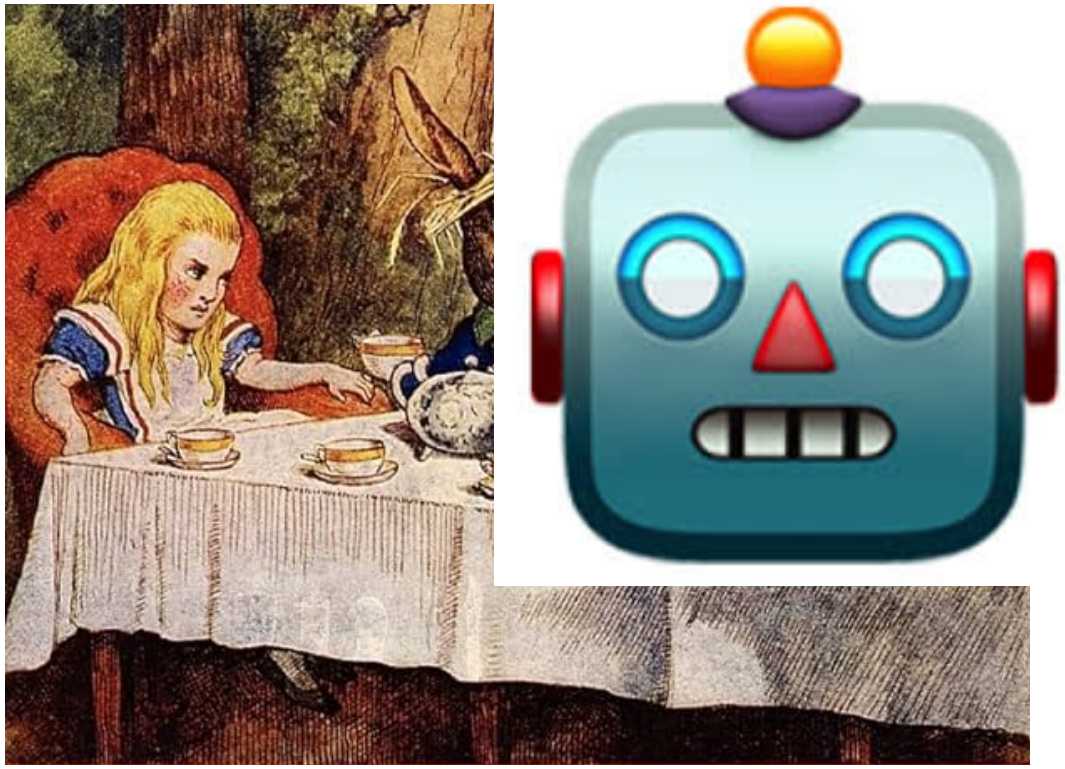
Our Framework

1. Requires a quite strong common prior btwn agents

2. *Humans* are unrealistically modeled as Bayesians

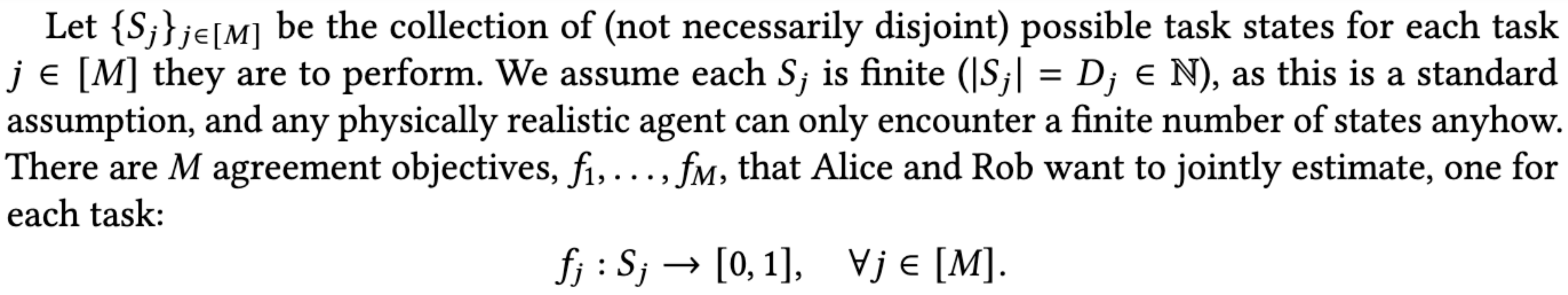
3. Does not study how long the coordination will take (in terms of number of messages)

Alice



“Rob”

Our Framework



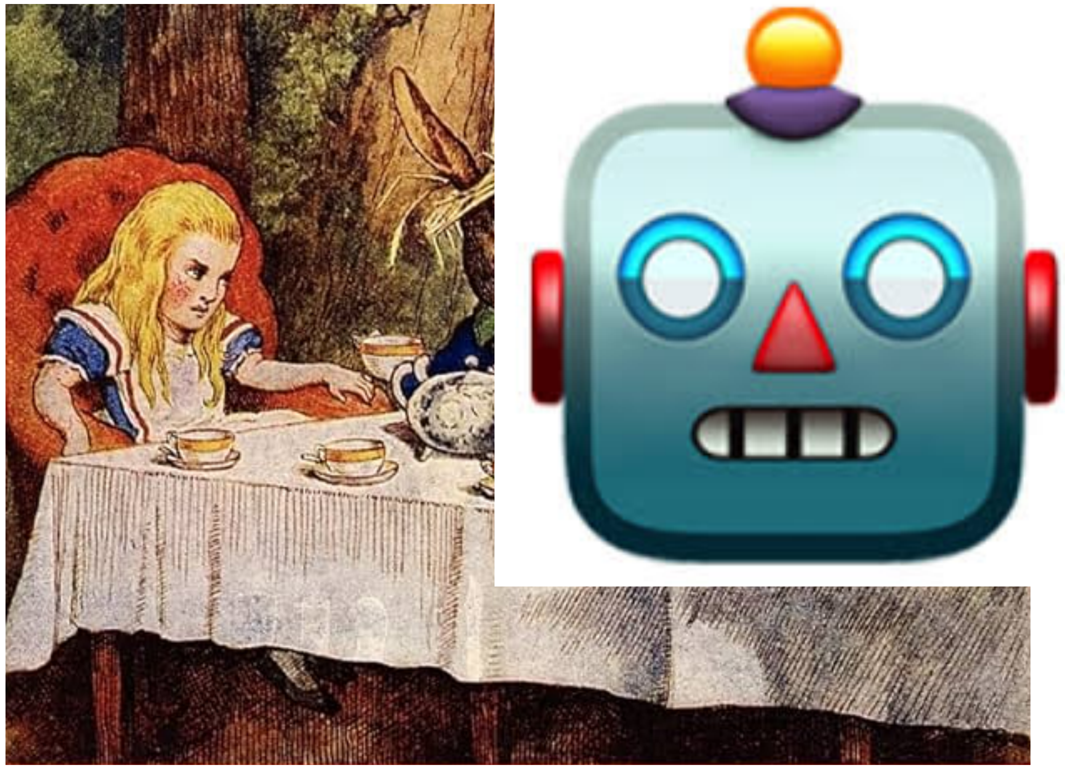
Let $\{S_j\}_{j\in[M]}$ be the collection of (not necessarily disjoint) possible task states for each task $j \in [M]$ they are to perform. We assume each S_j is finite ($|S_j| = D_j \in \mathbb{N}$), as this is a standard assumption, and any physically realistic agent can only encounter a finite number of states anyhow. There are M agreement objectives, $f_1, \ldots, f_M$, that Alice and Rob want to jointly estimate, one for each task:

$$f_j : S_j \to [0,1], \quad \forall j \in [M].$$

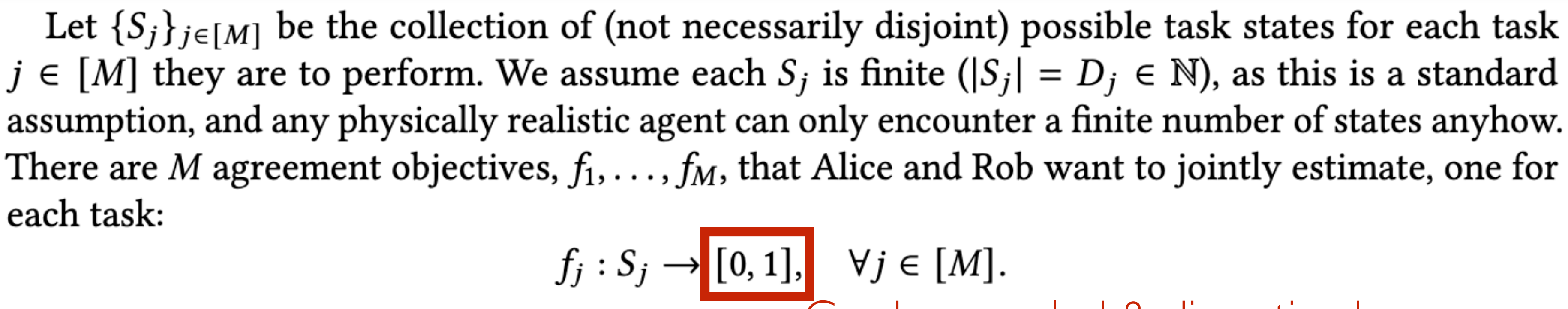
1. Requires a quite strong common prior btwn agents

2. *Humans* are unrealistically modeled as Bayesians

3. Does not study how long the coordination will take (in terms of number of messages)



Our Framework



Let $\{S_j\}_{j\in[M]}$ be the collection of (not necessarily disjoint) possible task states for each task $j \in [M]$ they are to perform. We assume each S_j is finite ($|S_j| = D_j \in \mathbb{N}$), as this is a standard assumption, and any physically realistic agent can only encounter a finite number of states anyhow. There are M agreement objectives, $f_1, \ldots, f_M$, that Alice and Rob want to jointly estimate, one for each task:

$$f_j : S_j \to \boxed{[0, 1],} \quad \forall j \in [M].$$

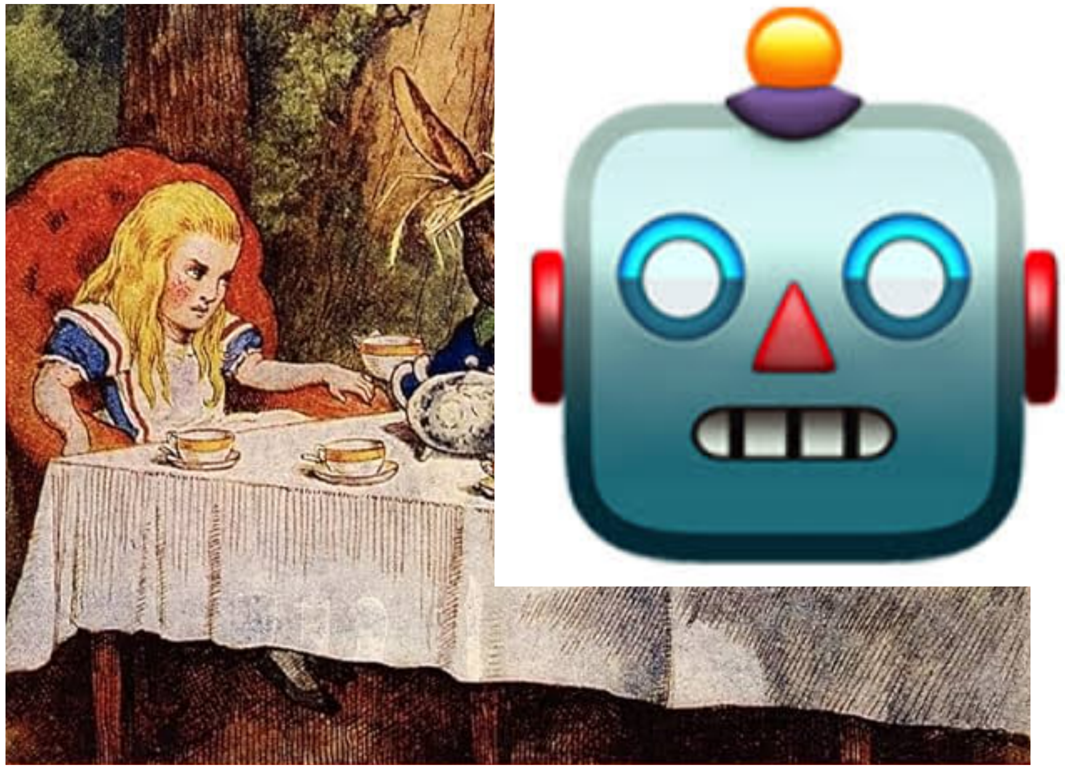
Can be rescaled & discretized

1. Requires a quite strong common prior btwn agents

2. *Humans* are unrealistically modeled as Bayesians

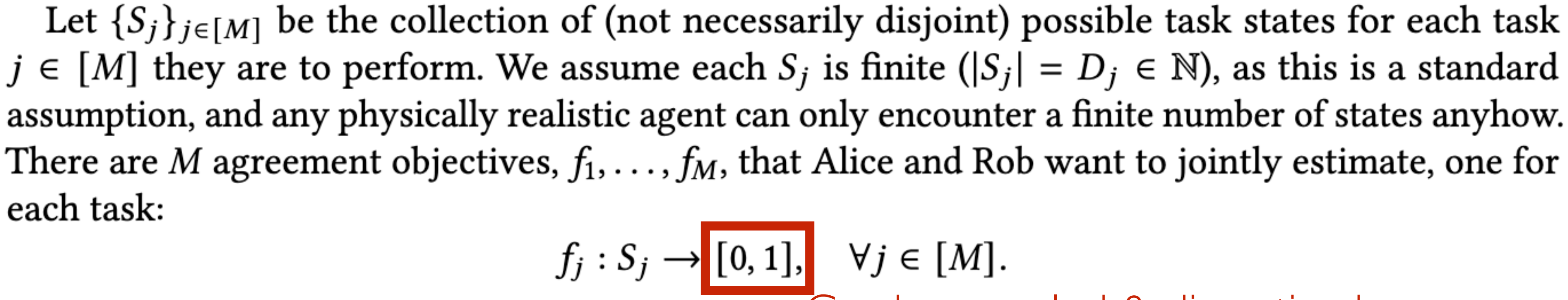
3. Does not study how long the coordination will take (in terms of number of messages)

Alice



"Rob"

Our Framework



Let $\{S_j\}_{j\in[M]}$ be the collection of (not necessarily disjoint) possible task states for each task $j \in [M]$ they are to perform. We assume each S_j is finite ($|S_j| = D_j \in \mathbb{N}$), as this is a standard assumption, and any physically realistic agent can only encounter a finite number of states anyhow. There are M agreement objectives, $f_1, \dots, f_M$, that Alice and Rob want to jointly estimate, one for each task:

$$f_j : S_j \to \boxed{[0,1],} \quad \forall j \in [M].$$

Can be rescaled & discretized

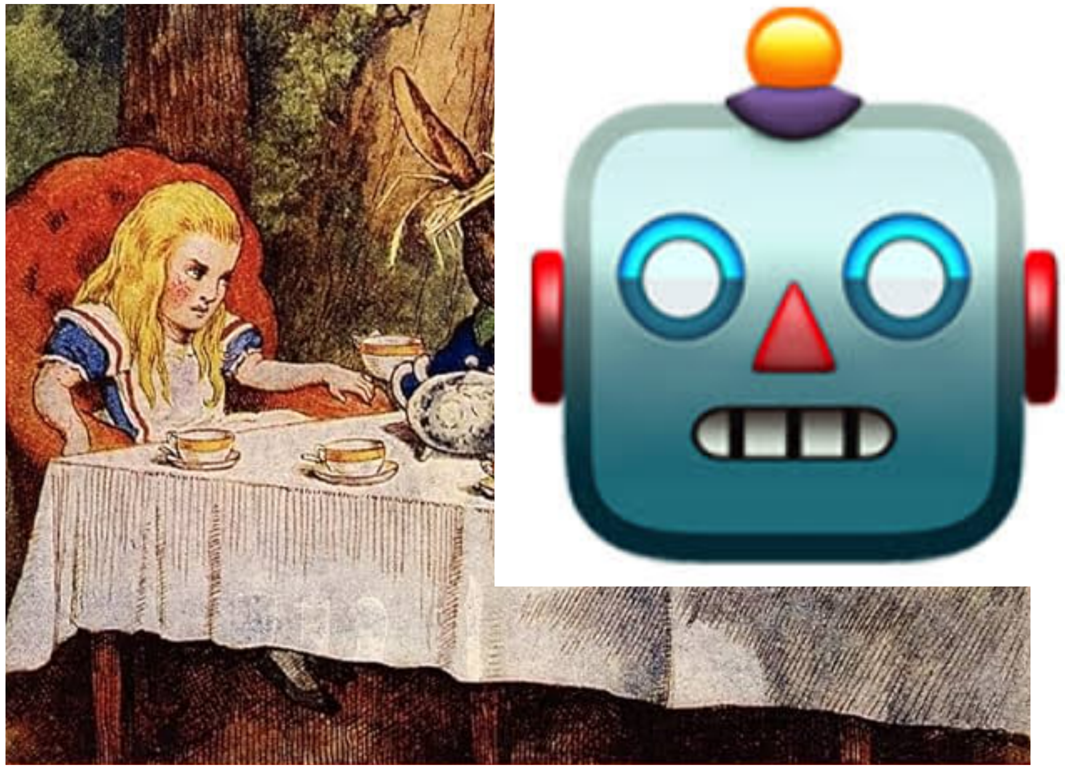
Exchange messages until: $m_j^1, \dots, m_j^T : \mathscr{P}(S_j) \to [0,1]$

1. Requires a quite strong common prior btwn agents

2. *Humans* are unrealistically modeled as Bayesians

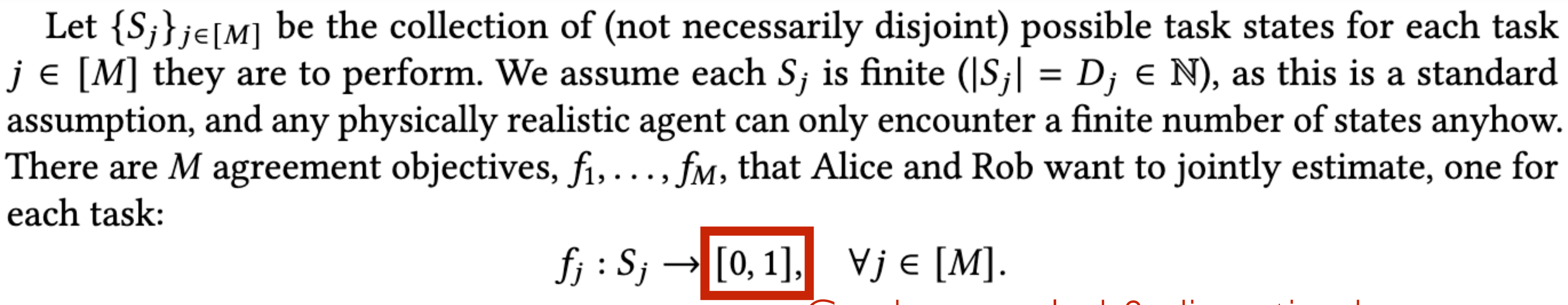
3. Does not study how long the coordination will take (in terms of number of messages)

Alice



"Rob"

Our Framework

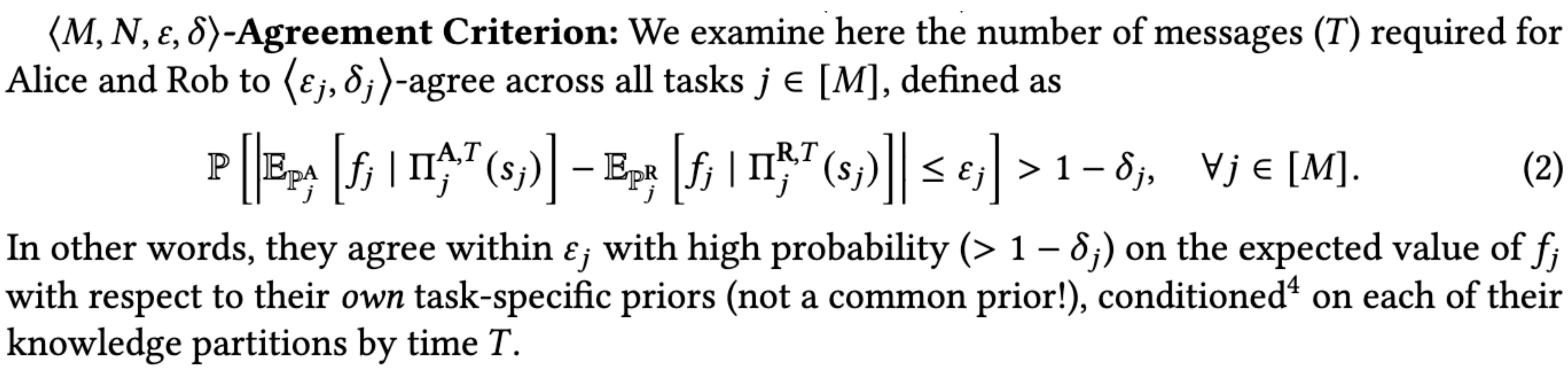


Let $\{S_j\}_{j\in[M]}$ be the collection of (not necessarily disjoint) possible task states for each task $j \in [M]$ they are to perform. We assume each S_j is finite ($|S_j| = D_j \in \mathbb{N}$), as this is a standard assumption, and any physically realistic agent can only encounter a finite number of states anyhow. There are M agreement objectives, $f_1, \ldots, f_M$, that Alice and Rob want to jointly estimate, one for each task:

$$f_j : S_j \to \boxed{[0,1],} \quad \forall j \in [M].$$

Can be rescaled & discretized

Exchange messages until: $m_j^1, \ldots, m_j^T : \mathscr{P}(S_j) \to [0,1]$

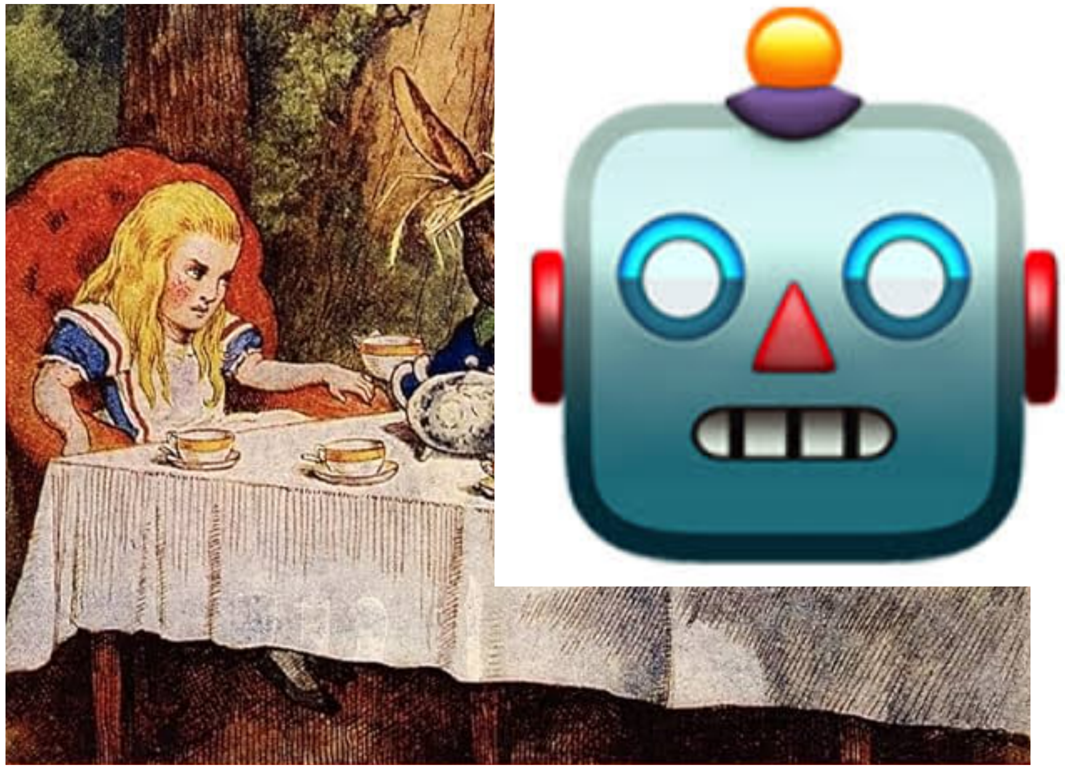


$\langle M, N, \varepsilon, \delta\rangle$-**Agreement Criterion:** We examine here the number of messages (T) required for Alice and Rob to $\langle \varepsilon_j, \delta_j\rangle$-agree across all tasks $j \in [M]$, defined as

$$\mathbb{P}\left[\left|\mathbb{E}_{\mathbb{P}_j^{\mathrm{A}}}\left[f_j \mid \Pi_j^{\mathrm{A},T}(s_j)\right] - \mathbb{E}_{\mathbb{P}_j^{\mathrm{R}}}\left[f_j \mid \Pi_j^{\mathrm{R},T}(s_j)\right]\right| \le \varepsilon_j\right] > 1-\delta_j, \quad \forall j \in [M]. \qquad (2)$$

In other words, they agree within ε_j with high probability ($> 1 - \delta_j$) on the expected value of f_j with respect to their *own* task-specific priors (not a common prior!), conditioned[4] on each of their knowledge partitions by time T.

Alice



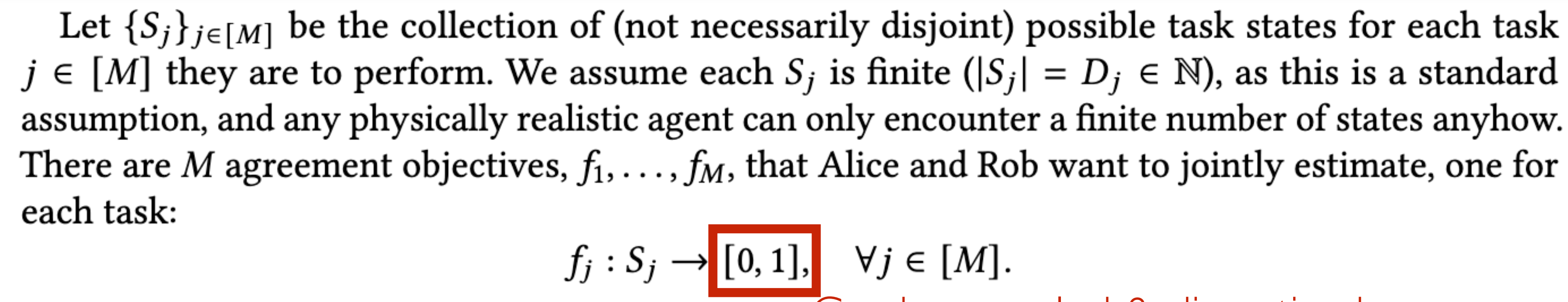
“Rob”

1. Requires a quite strong common prior btwn agents

2. *Humans* are unrealistically modeled as Bayesians

3. Does not study how long the coordination will take (in terms of number of messages)

Our Framework

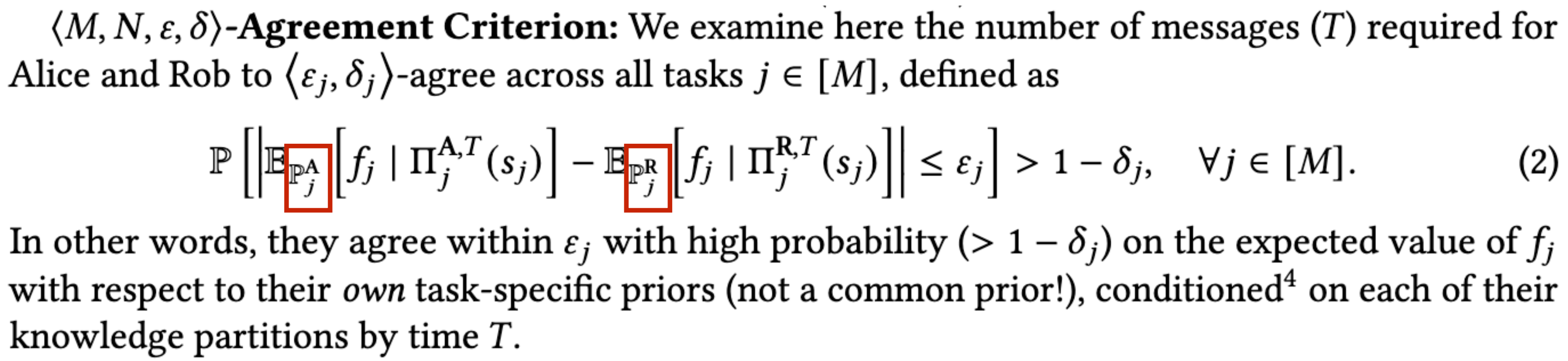


Let $\{S_j\}_{j\in[M]}$ be the collection of (not necessarily disjoint) possible task states for each task $j \in [M]$ they are to perform. We assume each S_j is finite ($|S_j| = D_j \in \mathbb{N}$), as this is a standard assumption, and any physically realistic agent can only encounter a finite number of states anyhow. There are M agreement objectives, $f_1, \ldots, f_M$, that Alice and Rob want to jointly estimate, one for each task:

$$f_j : S_j \to \boxed{[0, 1]}, \quad \forall j \in [M].$$

Can be rescaled & discretized

Exchange messages until: $m_j^1, \ldots, m_j^T : \mathscr{P}(S_j) \to [0, 1]$

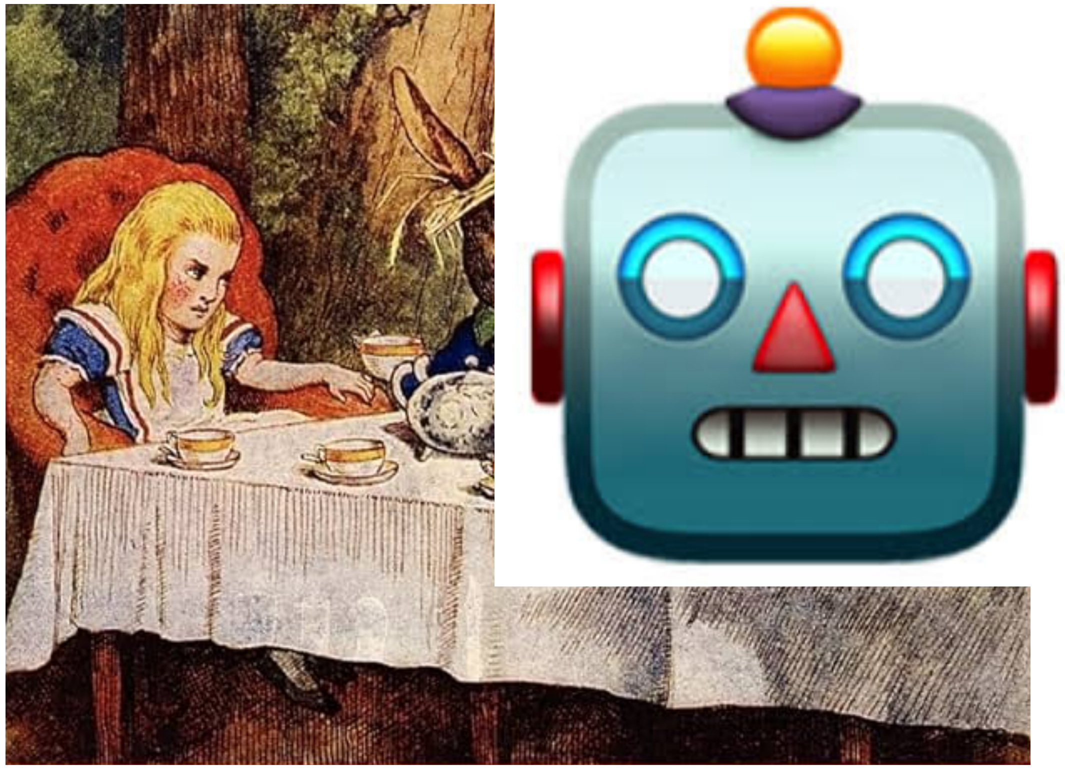


$\langle M, N, \varepsilon, \delta\rangle$-**Agreement Criterion:** We examine here the number of messages (T) required for Alice and Rob to $\langle \varepsilon_j, \delta_j\rangle$-agree across all tasks $j \in [M]$, defined as

$$\mathbb{P}\left[\left|\mathbb{E}_{\boxed{\mathbb{P}_j^{\mathrm{A}}}}\left[f_j \mid \Pi_j^{\mathrm{A},T}(s_j)\right] - \mathbb{E}_{\boxed{\mathbb{P}_j^{\mathrm{R}}}}\left[f_j \mid \Pi_j^{\mathrm{R},T}(s_j)\right]\right| \le \varepsilon_j\right] > 1 - \delta_j, \quad \forall j \in [M]. \tag{2}$$

In other words, they agree within ε_j with high probability ($> 1 - \delta_j$) on the expected value of f_j with respect to their *own* task-specific priors (not a common prior!), conditioned[4] on each of their knowledge partitions by time T.

Alice



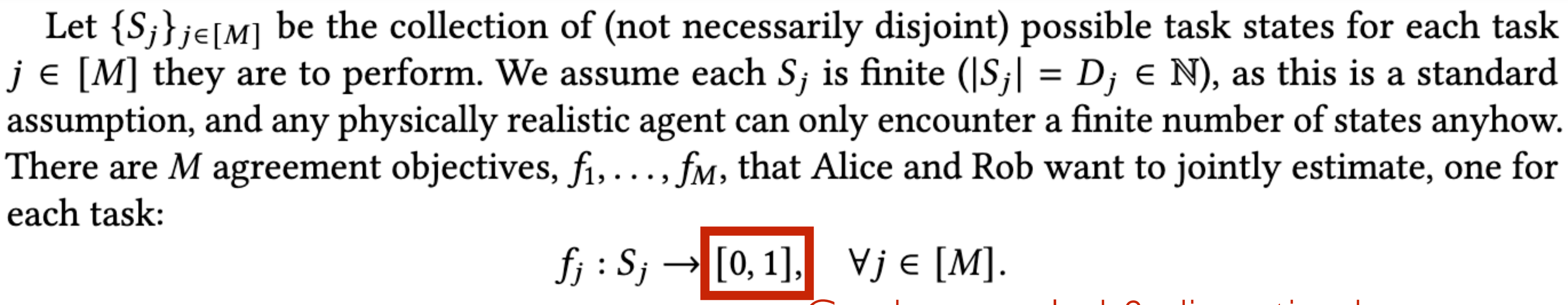
"Rob"

1. ~~Requires quite strong common priors between agents~~

2. *Humans* are unrealistically modeled as Bayesians

3. Does not study how long the coordination will take (in terms of number of messages)

Our Framework

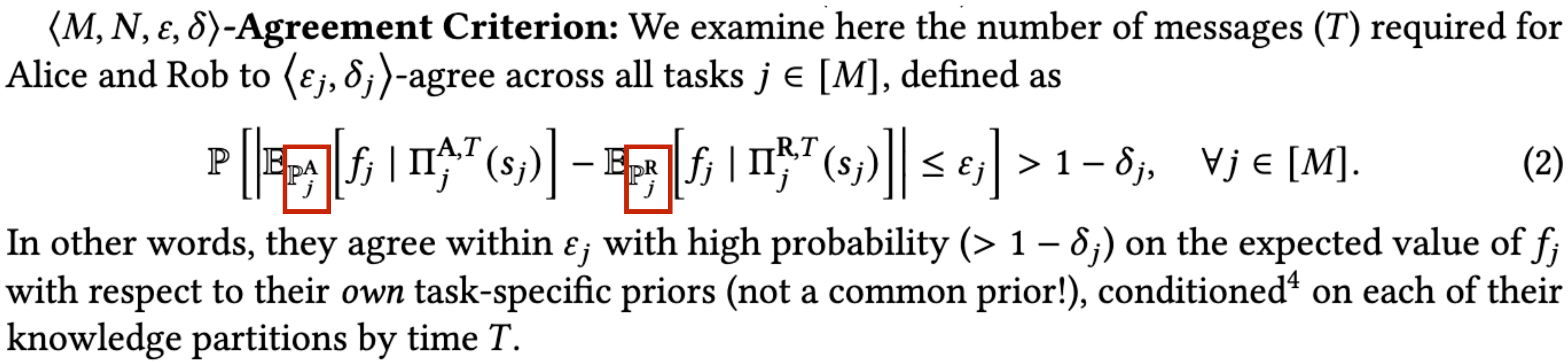


Let $\{S_j\}_{j\in[M]}$ be the collection of (not necessarily disjoint) possible task states for each task $j \in [M]$ they are to perform. We assume each S_j is finite ($|S_j| = D_j \in \mathbb{N}$), as this is a standard assumption, and any physically realistic agent can only encounter a finite number of states anyhow. There are M agreement objectives, $f_1, \ldots, f_M$, that Alice and Rob want to jointly estimate, one for each task:

$$f_j : S_j \to \boxed{[0,1],} \quad \forall j \in [M].$$

Can be rescaled & discretized

Exchange messages until: $m_j^1, \ldots, m_j^T : \mathscr{P}(S_j) \to [0,1]$

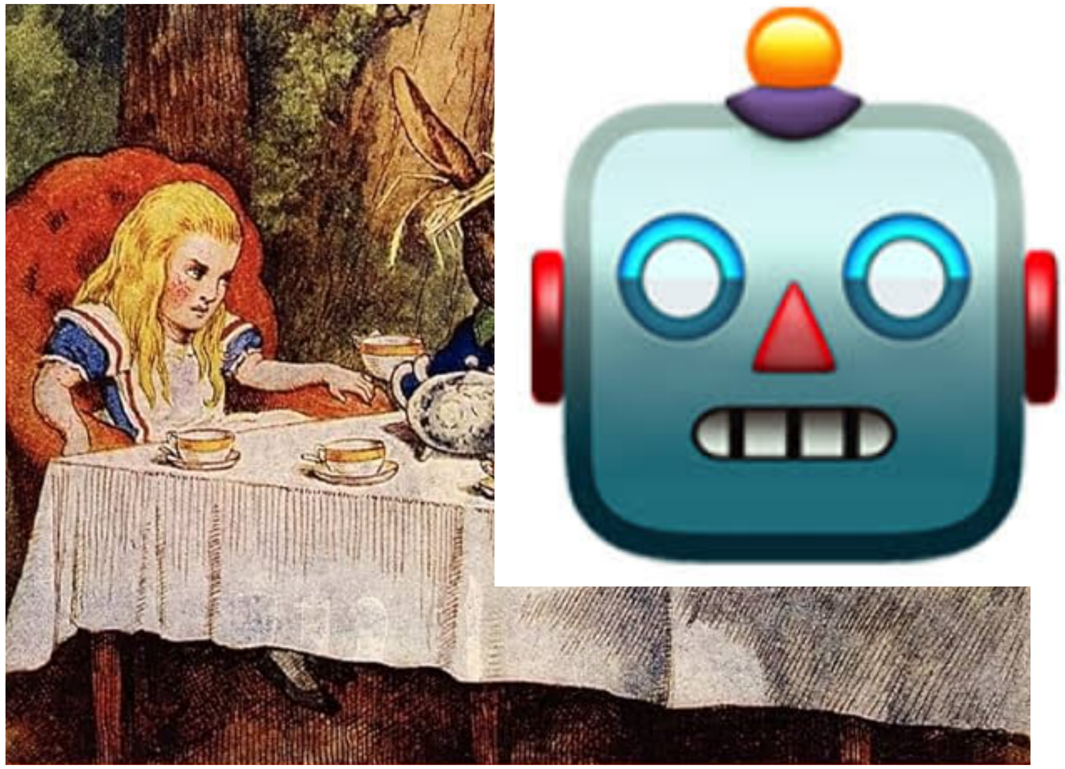


$\langle M, N, \varepsilon, \delta\rangle$-**Agreement Criterion:** We examine here the number of messages (T) required for Alice and Rob to $\langle \varepsilon_j, \delta_j\rangle$-agree across all tasks $j \in [M]$, defined as

$$\mathbb{P}\left[\left|\mathbb{E}_{\boxed{\mathbb{P}_j^{\mathrm{A}}}}\left[f_j \mid \Pi_j^{\mathrm{A},T}(s_j)\right] - \mathbb{E}_{\boxed{\mathbb{P}_j^{\mathrm{R}}}}\left[f_j \mid \Pi_j^{\mathrm{R},T}(s_j)\right]\right| \le \varepsilon_j\right] > 1 - \delta_j, \quad \forall j \in [M]. \tag{2}$$

In other words, they agree within ε_j with high probability ($> 1 - \delta_j$) on the expected value of f_j with respect to their *own* task-specific priors (not a common prior!), conditioned[4] on each of their knowledge partitions by time T.

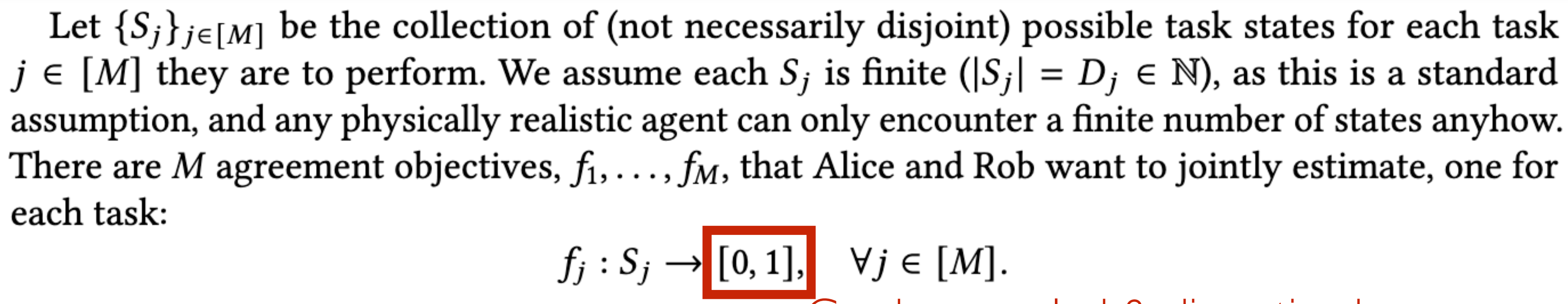
Alice



“Rob”

1. ~~Requires quite strong common priors over agents~~
2. ~~*Humans are* unrealistically modeled as Bayesians~~
3. Does not study how long the coordination will take (in terms of number of messages)

Our Framework

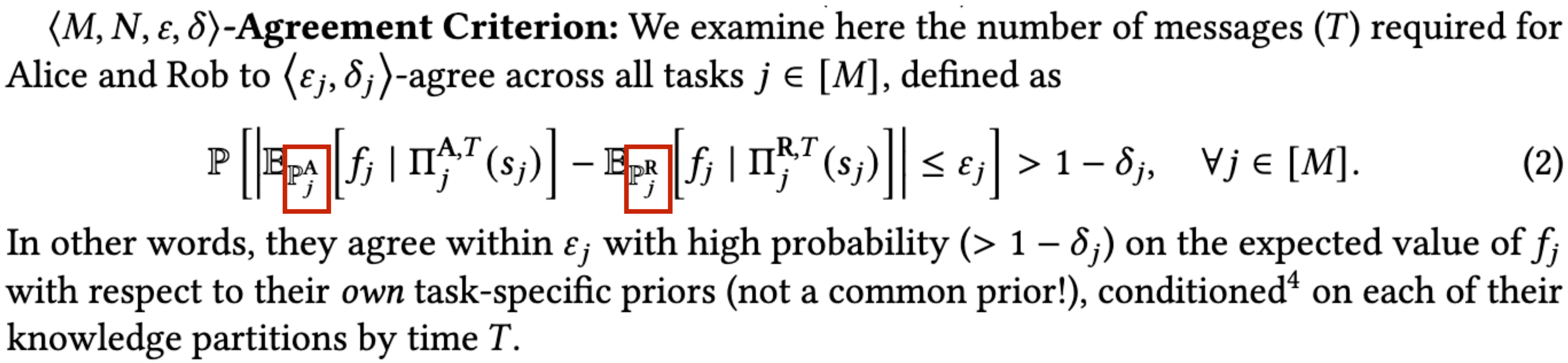


Let $\{S_j\}_{j\in[M]}$ be the collection of (not necessarily disjoint) possible task states for each task $j \in [M]$ they are to perform. We assume each S_j is finite ($|S_j| = D_j \in \mathbb{N}$), as this is a standard assumption, and any physically realistic agent can only encounter a finite number of states anyhow. There are M agreement objectives, $f_1, \ldots, f_M$, that Alice and Rob want to jointly estimate, one for each task:

$$f_j : S_j \to \boxed{[0,1]}, \quad \forall j \in [M].$$

Can be rescaled & discretized

Exchange messages until: $m_j^1, \ldots, m_j^T : \mathscr{P}(S_j) \to [0,1]$



$\langle M, N, \varepsilon, \delta\rangle$**-Agreement Criterion:** We examine here the number of messages (T) required for Alice and Rob to $\langle \varepsilon_j, \delta_j \rangle$-agree across all tasks $j \in [M]$, defined as

$$\mathbb{P}\left[\left|\mathbb{E}_{\boxed{\mathbb{P}_j^{\mathrm{A}}}}\left[f_j \mid \Pi_j^{\mathrm{A},T}(s_j)\right] - \mathbb{E}_{\boxed{\mathbb{P}_j^{\mathrm{R}}}}\left[f_j \mid \Pi_j^{\mathrm{R},T}(s_j)\right]\right| \le \varepsilon_j\right] > 1 - \delta_j, \quad \forall j \in [M]. \tag{2}$$

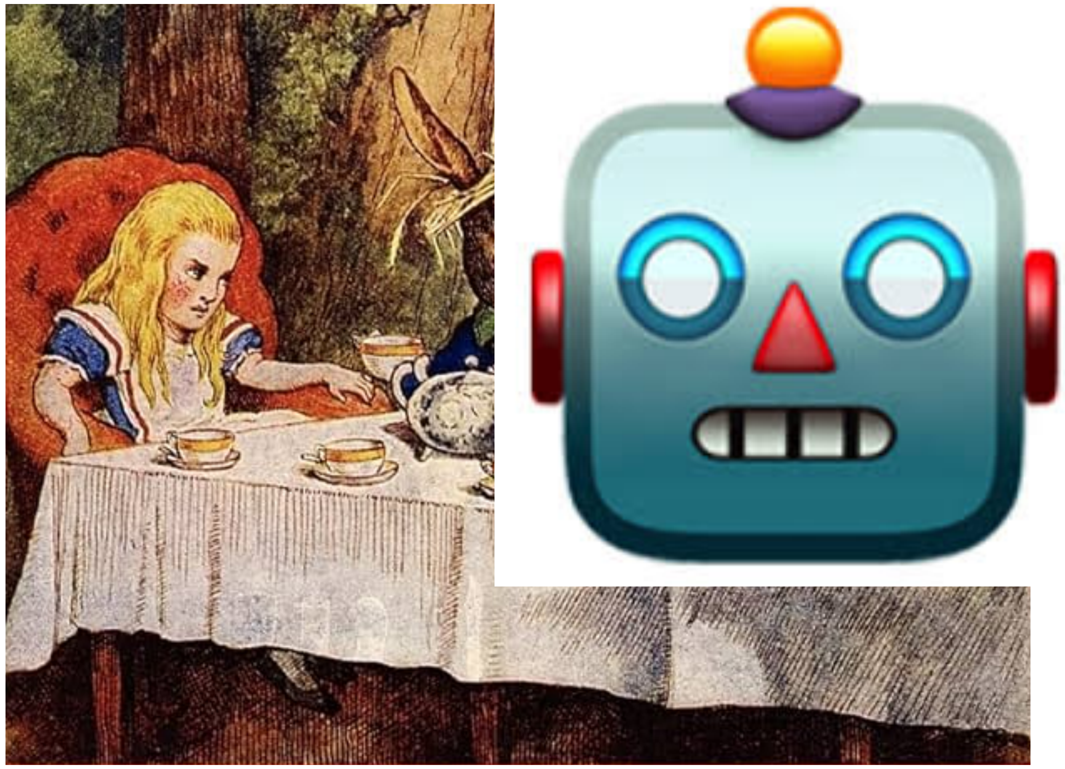
In other words, they agree within ε_j with high probability ($> 1 - \delta_j$) on the expected value of f_j with respect to their *own* task-specific priors (not a common prior!), conditioned[4] on each of their knowledge partitions by time T.

1. ~~Requires quite strong common priors for agents~~

2. ~~Humans are unrealistically modeled as Bayesians~~

3. Does not study how long the coordination will take (in terms of number of messages)

Alice



"Rob"

We study the communication complexity (# of messages) T & without requiring *exact* agreement

Our Framework

Let $\{S_j\}_{j \in [M]}$ be the collection of (not necessarily disjoint) possible task states for each task $j \in [M]$ they are to perform. We assume each S_j is finite ($|S_j| = D_j \in \mathbb{N}$), as this is a standard assumption, and any physically realistic agent can only encounter a finite number of states anyhow. There are M agreement objectives, $f_1, \ldots, f_M$, that Alice and Rob want to jointly estimate, one for each task:

$$f_j : S_j \to \boxed{[0, 1]}, \quad \forall j \in [M].$$

Can be rescaled & discretized

Exchange messages until: $m_j^1, \ldots, m_j^T : \mathscr{P}(S_j) \to [0, 1]$

$\langle M, N, \varepsilon, \delta \rangle$**-Agreement Criterion:** We examine here the number of messages (T) required for Alice and Rob to $\langle \varepsilon_j, \delta_j \rangle$-agree across all tasks $j \in [M]$, defined as

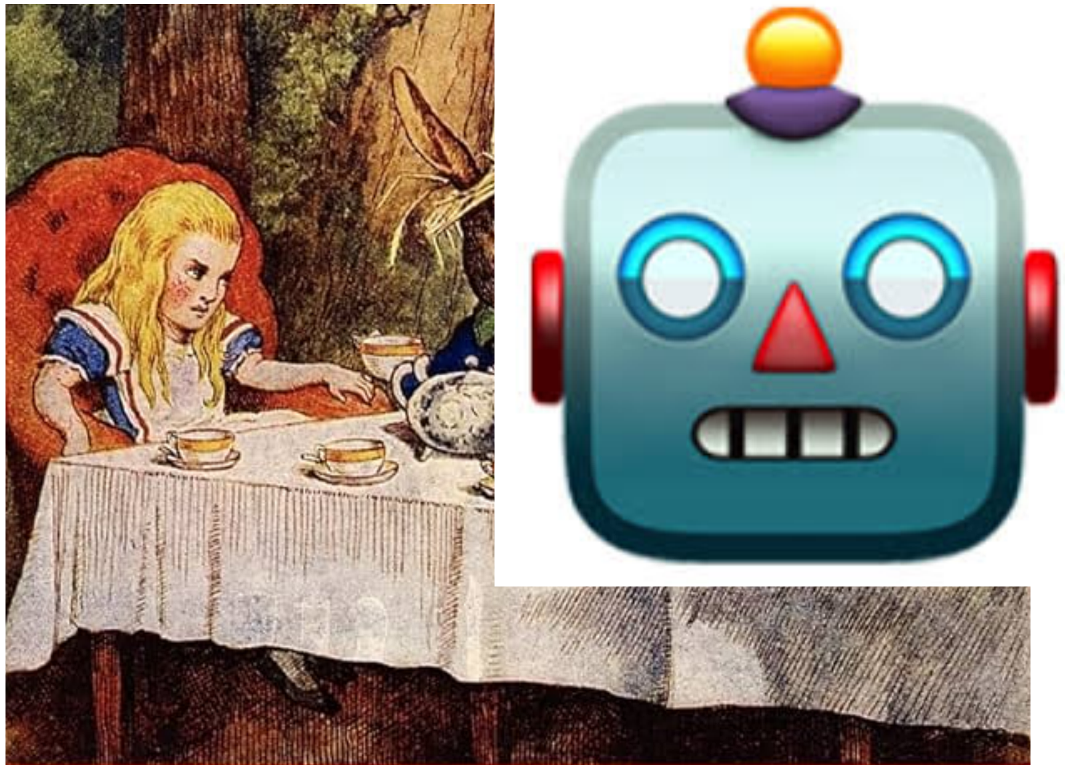
$$\mathbb{P}\left[\left|\mathbb{E}_{\boxed{\mathbb{P}_j^{\mathrm{A}}}}\left[f_j \mid \Pi_j^{\mathrm{A},T}(s_j)\right] - \mathbb{E}_{\boxed{\mathbb{P}_j^{\mathrm{R}}}}\left[f_j \mid \Pi_j^{\mathrm{R},T}(s_j)\right]\right| \leq \varepsilon_j\right] > 1 - \delta_j, \quad \forall j \in [M]. \tag{2}$$

In other words, they agree within ε_j with high probability ($> 1 - \delta_j$) on the expected value of f_j with respect to their *own* task-specific priors (not a common prior!), conditioned[4] on each of their knowledge partitions by time T.

1. ~~Requires quite strong common priors between agents~~
2. ~~*Humans* are unrealistically modeled as Bayesians~~
3. ~~Does not study how long the coordination will take (in terms of number of messages)~~

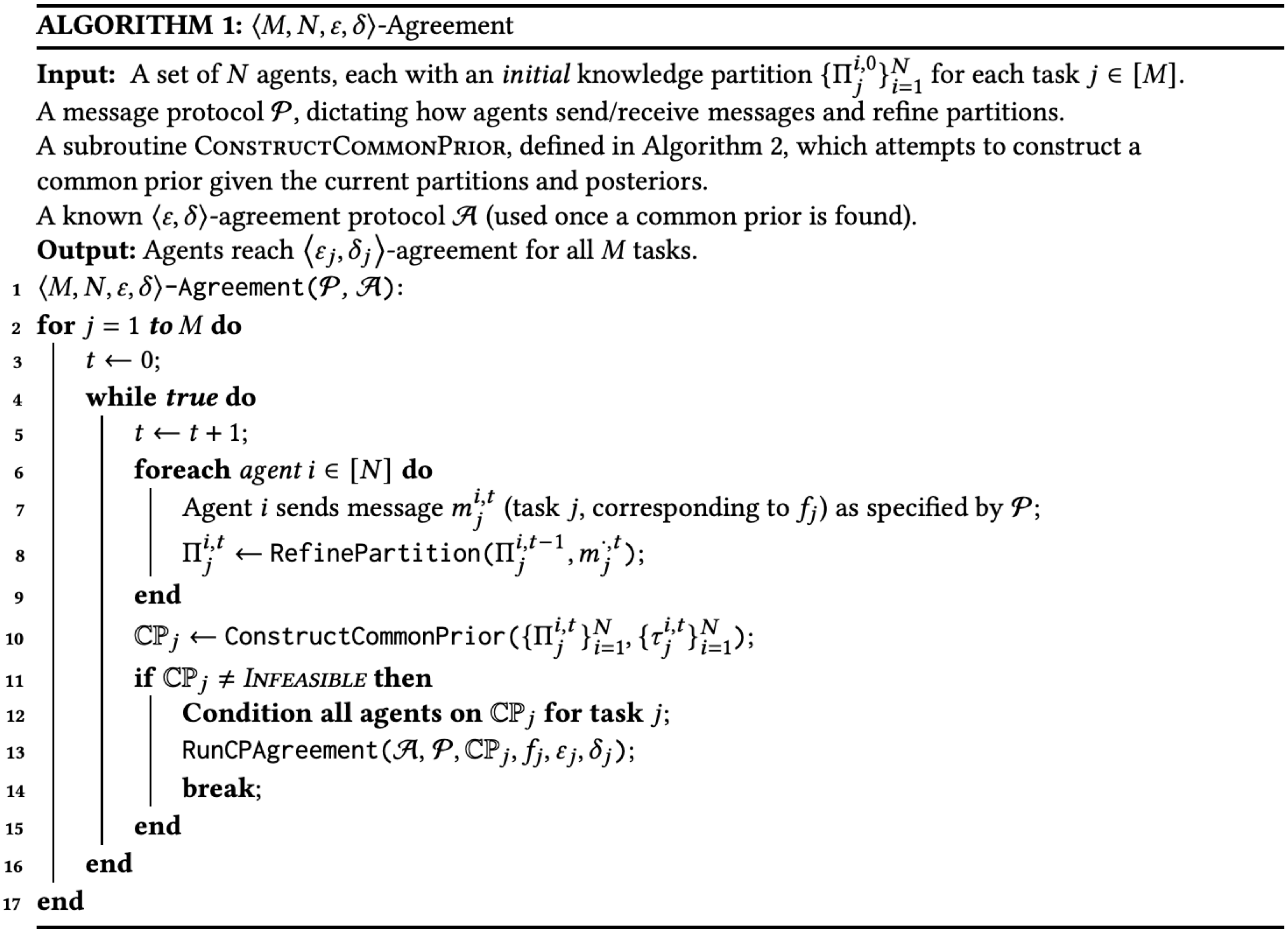
We study the communication complexity (# of messages) T & without requiring *exact* agreement

Alice



"Rob"

Our Framework: Explicit Algorithm

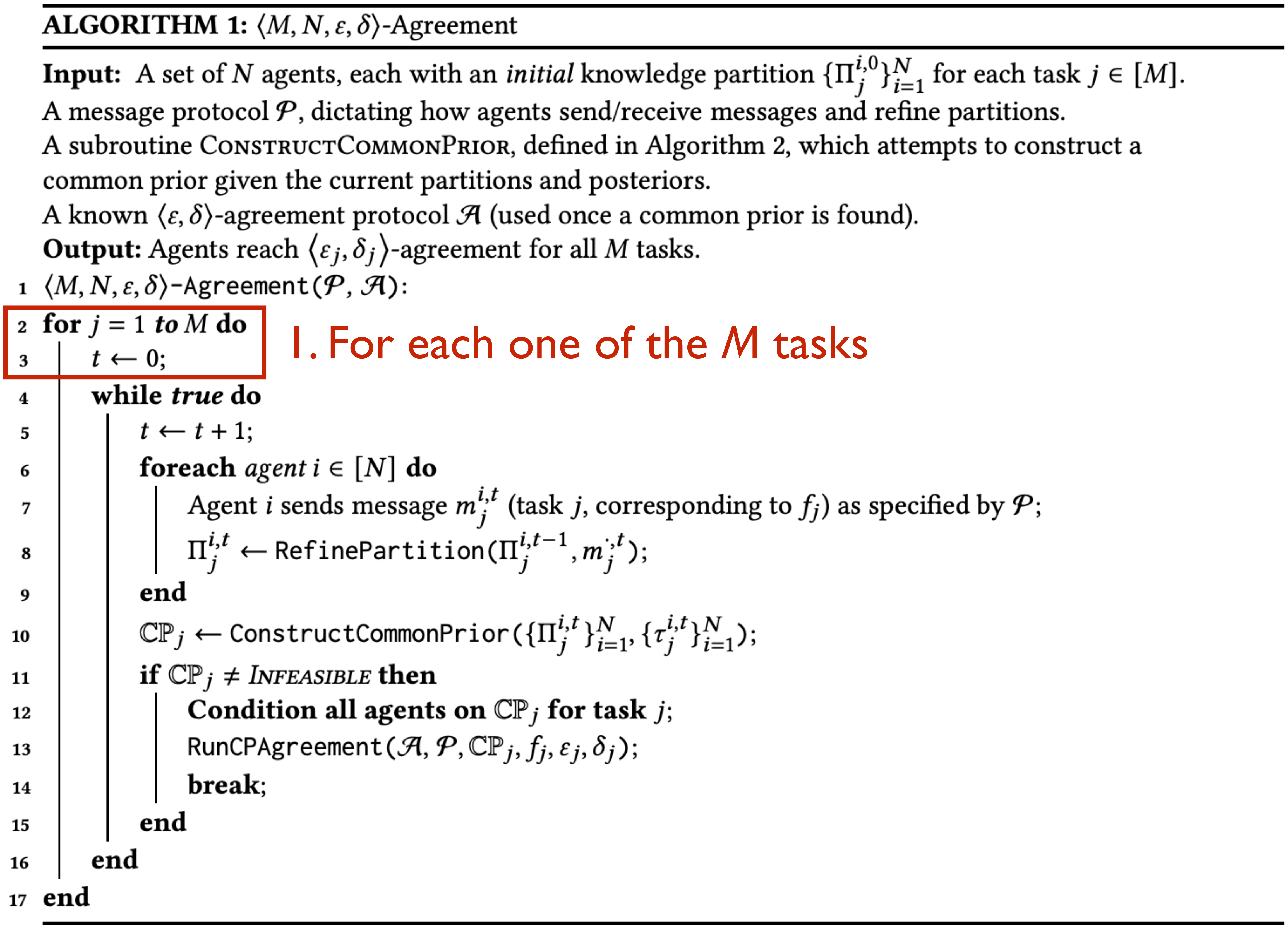


ALGORITHM 1: $\langle M, N, \varepsilon, \delta \rangle$-Agreement

Input: A set of N agents, each with an *initial* knowledge partition $\{\Pi_j^{i,0}\}_{i=1}^N$ for each task $j \in [M]$.
A message protocol $\mathcal{P}$, dictating how agents send/receive messages and refine partitions.
A subroutine ConstructCommonPrior, defined in Algorithm 2, which attempts to construct a common prior given the current partitions and posteriors.
A known $\langle \varepsilon, \delta \rangle$-agreement protocol $\mathcal{A}$ (used once a common prior is found).
Output: Agents reach $\langle \varepsilon_j, \delta_j \rangle$-agreement for all M tasks.

```
⟨M, N, ε, δ⟩-Agreement(𝒫, 𝒜):
for $j = 1$ to $M$ do
    $t \leftarrow 0$;
    while true do
        $t \leftarrow t + 1$;
        foreach agent $i \in [N]$ do
            Agent $i$ sends message $m_j^{i,t}$ (task $j$, corresponding to $f_j$) as specified by $\mathcal{P}$;
            $\Pi_j^{i,t} \leftarrow$ RefinePartition($\Pi_j^{i,t-1}, m_j^{\cdot,t}$);
        end
        $\mathbb{CP}_j \leftarrow$ ConstructCommonPrior($\{\Pi_j^{i,t}\}_{i=1}^N, \{\tau_j^{i,t}\}_{i=1}^N$);
        if $\mathbb{CP}_j \neq$ Infeasible then
            Condition all agents on $\mathbb{CP}_j$ for task $j$;
            RunCPAgreement($\mathcal{A}, \mathcal{P}, \mathbb{CP}_j, f_j, \varepsilon_j, \delta_j$);
            break;
        end
    end
end
```

Our Framework: Explicit Algorithm



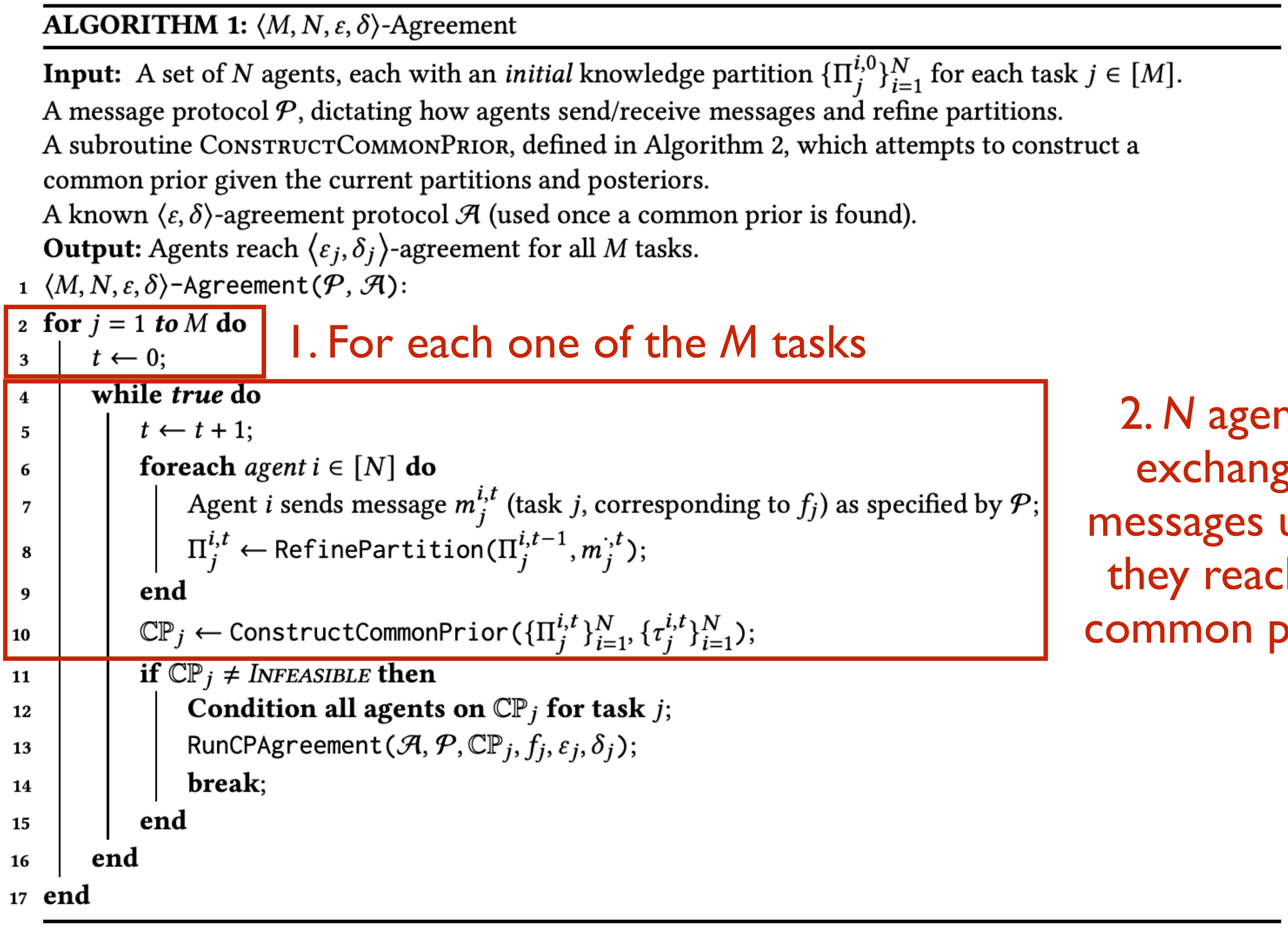
ALGORITHM 1: $\langle M, N, \varepsilon, \delta \rangle$-Agreement

Input: A set of N agents, each with an *initial* knowledge partition $\{\Pi_j^{i,0}\}_{i=1}^N$ for each task $j \in [M]$.
A message protocol $\mathcal{P}$, dictating how agents send/receive messages and refine partitions.
A subroutine ConstructCommonPrior, defined in Algorithm 2, which attempts to construct a common prior given the current partitions and posteriors.
A known $\langle \varepsilon, \delta \rangle$-agreement protocol $\mathcal{A}$ (used once a common prior is found).
Output: Agents reach $\langle \varepsilon_j, \delta_j \rangle$-agreement for all M tasks.

```
⟨M, N, ε, δ⟩-Agreement(𝒫, 𝒜):
for $j = 1$ to $M$ do
    $t \leftarrow 0$;
    while true do
        $t \leftarrow t + 1$;
        foreach agent $i \in [N]$ do
            Agent $i$ sends message $m_j^{i,t}$ (task $j$, corresponding to $f_j$) as specified by $\mathcal{P}$;
            $\Pi_j^{i,t} \leftarrow$ RefinePartition($\Pi_j^{i,t-1}, m_j^{\cdot,t}$);
        end
        $\mathbb{CP}_j \leftarrow$ ConstructCommonPrior($\{\Pi_j^{i,t}\}_{i=1}^N, \{\tau_j^{i,t}\}_{i=1}^N$);
        if $\mathbb{CP}_j \neq$ Infeasible then
            Condition all agents on $\mathbb{CP}_j$ for task $j$;
            RunCPAgreement($\mathcal{A}, \mathcal{P}, \mathbb{CP}_j, f_j, \varepsilon_j, \delta_j$);
            break;
        end
    end
end
```

1. For each one of the *M* tasks

Our Framework: Explicit Algorithm



ALGORITHM 1: $\langle M, N, \varepsilon, \delta\rangle$-Agreement

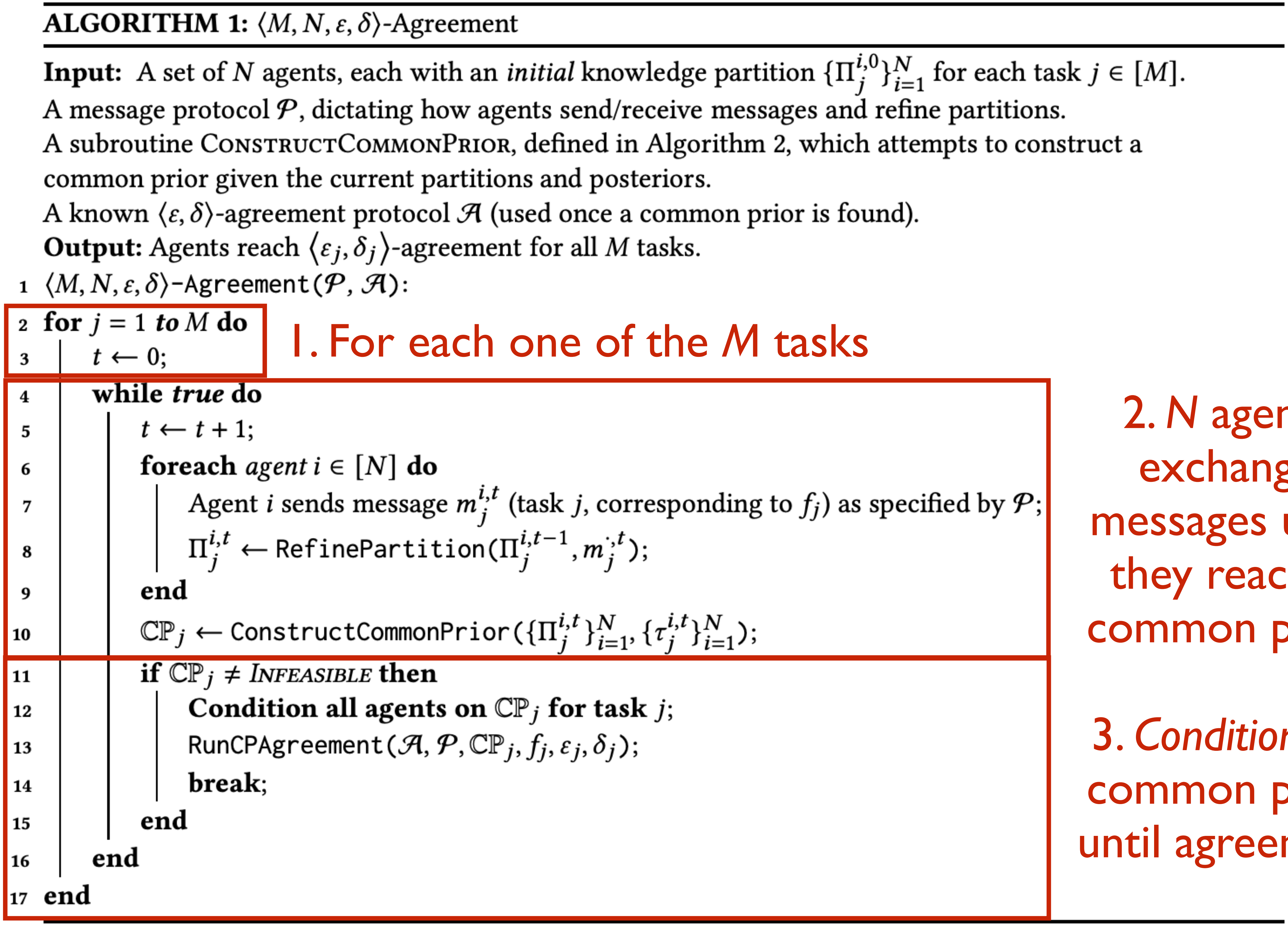
Input: A set of N agents, each with an *initial* knowledge partition $\{\Pi_j^{i,0}\}_{i=1}^N$ for each task $j \in [M]$.
A message protocol $\mathcal{P}$, dictating how agents send/receive messages and refine partitions.
A subroutine ConstructCommonPrior, defined in Algorithm 2, which attempts to construct a common prior given the current partitions and posteriors.
A known $\langle\varepsilon, \delta\rangle$-agreement protocol $\mathcal{A}$ (used once a common prior is found).
Output: Agents reach $\langle\varepsilon_j, \delta_j\rangle$-agreement for all M tasks.

```
⟨M, N, ε, δ⟩-Agreement(𝒫, 𝒜):
for $j = 1$ to $M$ do
    $t \leftarrow 0$;
    while true do
        $t \leftarrow t + 1$;
        foreach agent $i \in [N]$ do
            Agent $i$ sends message $m_j^{i,t}$ (task $j$, corresponding to $f_j$) as specified by $\mathcal{P}$;
            $\Pi_j^{i,t} \leftarrow$ RefinePartition($\Pi_j^{i,t-1}, m_j^{\cdot,t}$);
        end
        $\mathbb{CP}_j \leftarrow$ ConstructCommonPrior($\{\Pi_j^{i,t}\}_{i=1}^N, \{\tau_j^{i,t}\}_{i=1}^N$);
        if $\mathbb{CP}_j \neq$ INFEASIBLE then
            Condition all agents on $\mathbb{CP}_j$ for task $j$;
            RunCPAgreement($\mathcal{A}, \mathcal{P}, \mathbb{CP}_j, f_j, \varepsilon_j, \delta_j$);
            break;
        end
    end
end
```

1. For each one of the *M* tasks

2. *N* agents exchange messages until they reach a common prior

Our Framework: Explicit Algorithm



ALGORITHM 1: $\langle M, N, \varepsilon, \delta \rangle$-Agreement

Input: A set of N agents, each with an *initial* knowledge partition $\{\Pi_j^{i,0}\}_{i=1}^N$ for each task $j \in [M]$.
A message protocol $\mathcal{P}$, dictating how agents send/receive messages and refine partitions.
A subroutine ConstructCommonPrior, defined in Algorithm 2, which attempts to construct a common prior given the current partitions and posteriors.
A known $\langle \varepsilon, \delta \rangle$-agreement protocol $\mathcal{A}$ (used once a common prior is found).
Output: Agents reach $\langle \varepsilon_j, \delta_j \rangle$-agreement for all M tasks.

```
⟨M, N, ε, δ⟩-Agreement(P, A):
for j = 1 to M do
    t ← 0;
    while true do
        t ← t + 1;
        foreach agent i ∈ [N] do
            Agent i sends message $m_j^{i,t}$ (task j, corresponding to $f_j$) as specified by P;
            $\Pi_j^{i,t}$ ← RefinePartition($\Pi_j^{i,t-1}$, $m_j^{\cdot,t}$);
        end
        $\mathbb{CP}_j$ ← ConstructCommonPrior($\{\Pi_j^{i,t}\}_{i=1}^N$, $\{\tau_j^{i,t}\}_{i=1}^N$);
        if $\mathbb{CP}_j \neq$ Infeasible then
            Condition all agents on $\mathbb{CP}_j$ for task j;
            RunCPAgreement(A, P, $\mathbb{CP}_j$, $f_j$, $\varepsilon_j$, $\delta_j$);
            break;
        end
    end
end
```

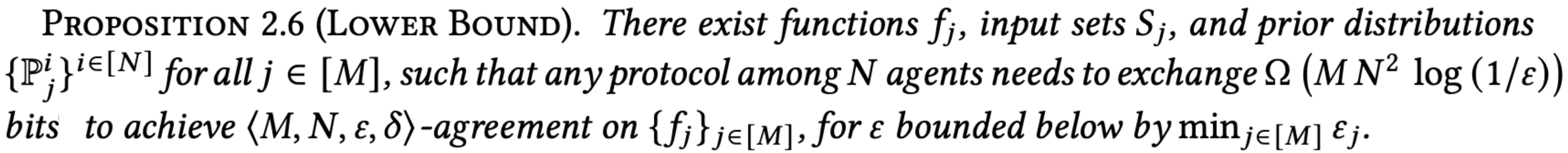
1. For each one of the *M* tasks

2. *N* agents exchange messages until they reach a common prior

3. *Condition* on common prior until agreement

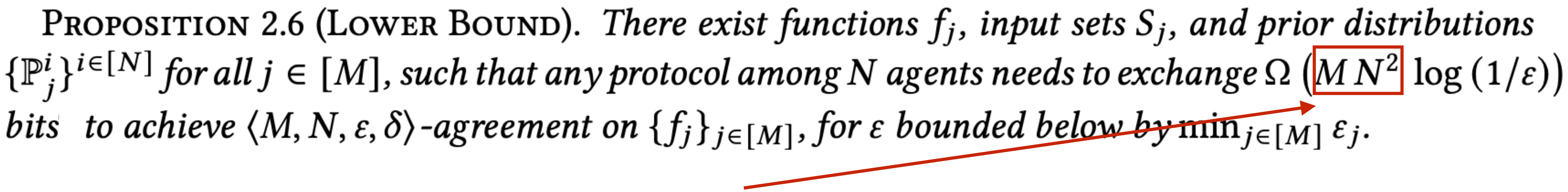
Fundamental Lower Bound: Unbounded Agent Setting

Fundamental Lower Bound: Unbounded Agent Setting



Proposition 2.6 (Lower Bound). *There exist functions f_j, input sets S_j, and prior distributions $\{\mathbb{P}^i_j\}^{i\in[N]}$ for all $j \in [M]$, such that any protocol among N agents needs to exchange $\Omega\left(M N^2 \log(1/\varepsilon)\right)$ bits to achieve $\langle M, N, \varepsilon, \delta\rangle$-agreement on $\{f_j\}_{j\in[M]}$, for ε bounded below by $\min_{j\in[M]} \varepsilon_j$.*

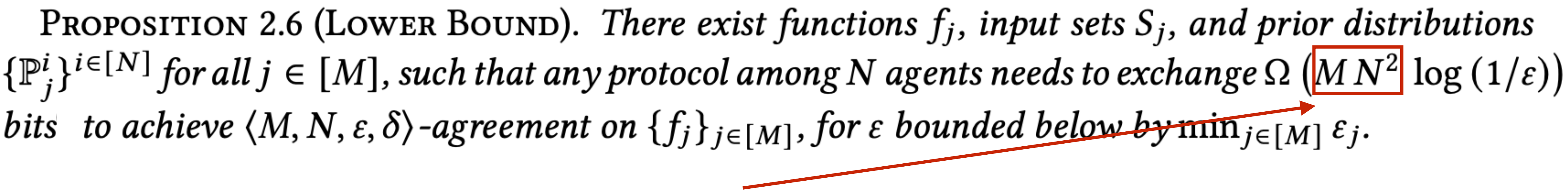
Fundamental Lower Bound: Unbounded Agent Setting



Proposition 2.6 (Lower Bound). *There exist functions f_j, input sets S_j, and prior distributions $\{\mathbb{P}^i_j\}^{i\in[N]}$ for all $j \in [M]$, such that any protocol among N agents needs to exchange $\Omega\left(\boxed{M N^2} \log(1/\varepsilon)\right)$ bits to achieve $\langle M, N, \varepsilon, \delta\rangle$-agreement on $\{f_j\}_{j\in[M]}$, for ε bounded below by $\min_{j\in[M]} \varepsilon_j$.*

If we have a large number of tasks (*M*) or agents (*N*), then it is *impossible* to *always* align them efficiently, even if the agents are computationally unbounded.

Fundamental Lower Bound: Unbounded Agent Setting

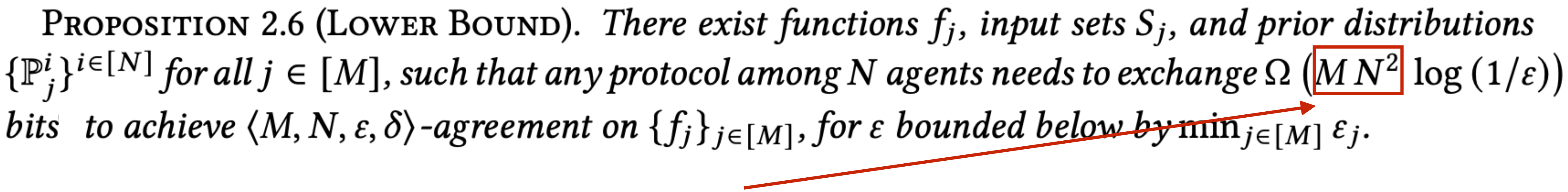


Proposition 2.6 (Lower Bound). *There exist functions f_j, input sets S_j, and prior distributions $\{\mathbb{P}^i_j\}^{i\in[N]}$ for all $j \in [M]$, such that any protocol among N agents needs to exchange $\Omega\left(\boxed{M N^2} \log(1/\varepsilon)\right)$ bits to achieve $\langle M, N, \varepsilon, \delta\rangle$-agreement on $\{f_j\}_{j\in[M]}$, for ε bounded below by $\min_{j\in[M]} \varepsilon_j$.*

If we have a large number of tasks (*M*) or agents (*N*), then it is *impossible* to *always* align them efficiently, even if the agents are computationally unbounded.

We need to choose our tasks & agents wisely!

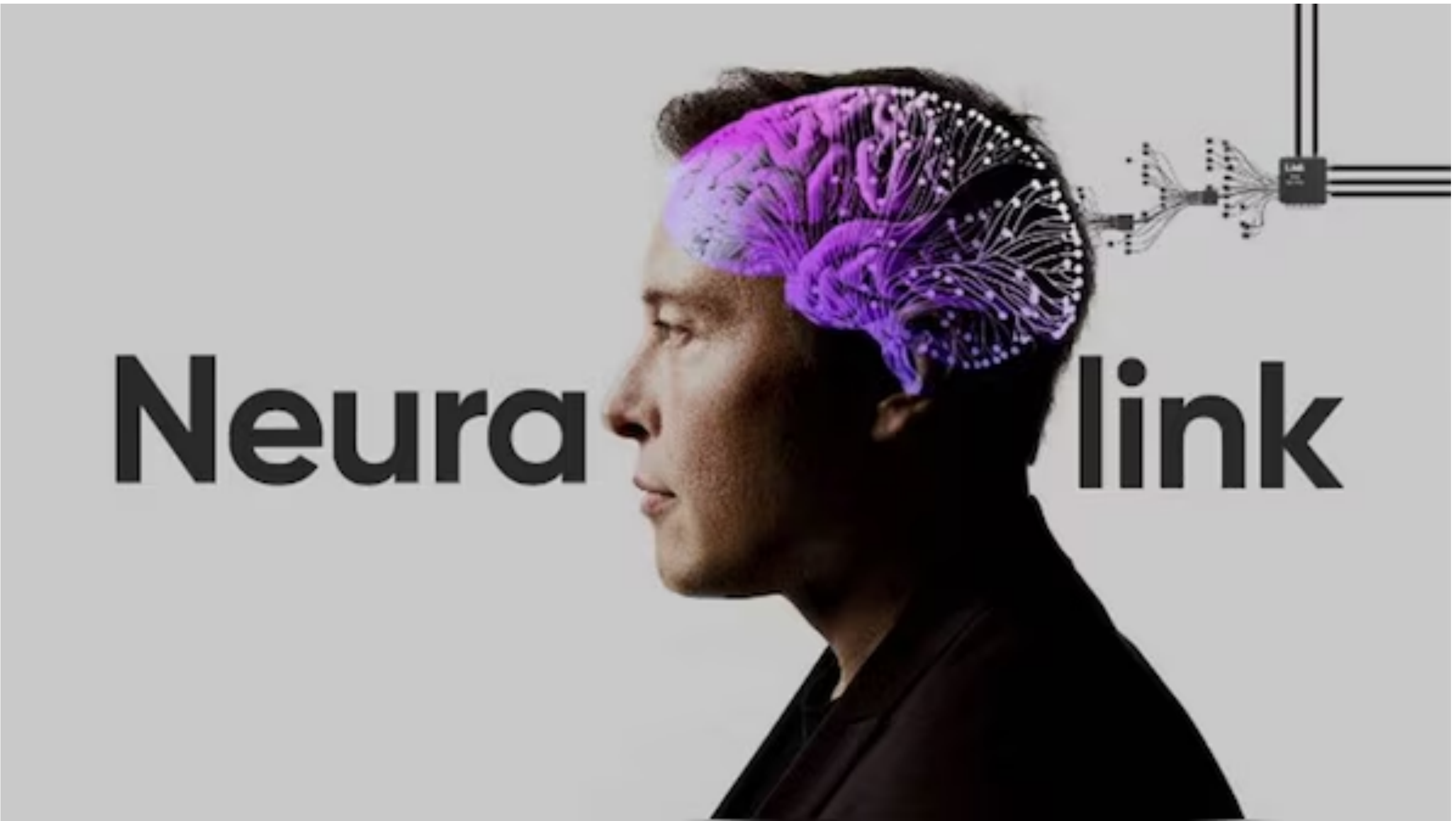
Fundamental Lower Bound: Unbounded Agent Setting



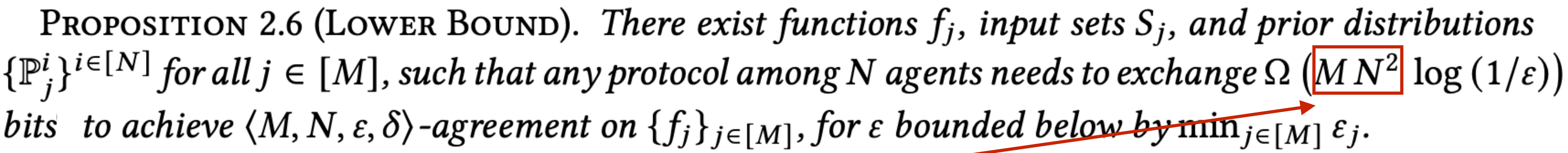
Proposition 2.6 (Lower Bound). *There exist functions f_j, input sets S_j, and prior distributions $\{\mathbb{P}^i_j\}^{i\in[N]}$ for all $j \in [M]$, such that any protocol among N agents needs to exchange $\Omega\left(\boxed{M N^2}\log(1/\varepsilon)\right)$ bits to achieve $\langle M, N, \varepsilon, \delta\rangle$-agreement on $\{f_j\}_{j\in[M]}$, for ε bounded below by $\min_{j\in[M]} \varepsilon_j$.*

If we have a large number of tasks (*M*) or agents (*N*), then it is *impossible* to *always* align them efficiently, even if the agents are computationally unbounded.

We need to choose our tasks & agents wisely!



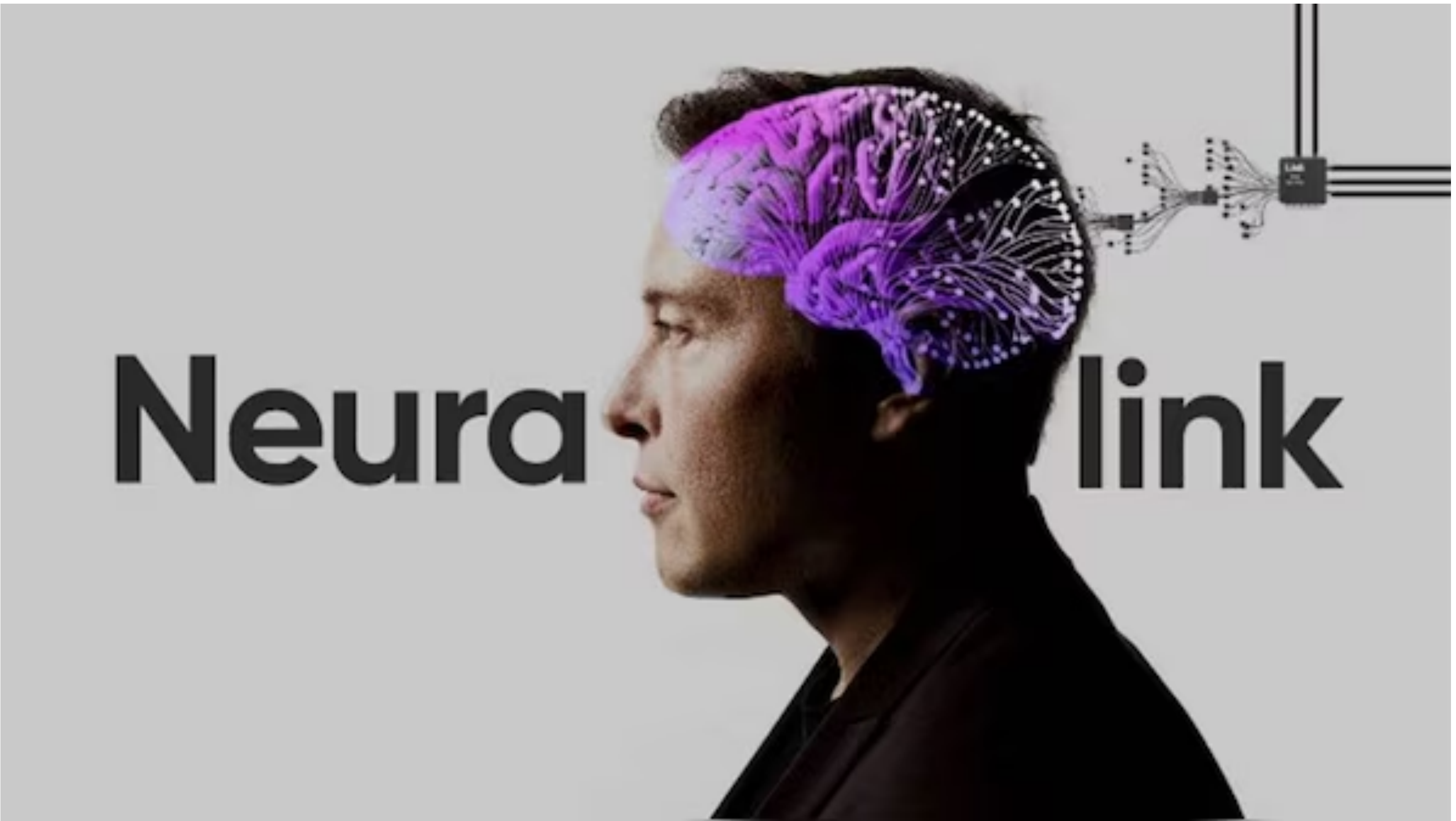
Fundamental Lower Bound: Unbounded Agent Setting



Proposition 2.6 (Lower Bound). *There exist functions f_j, input sets S_j, and prior distributions $\{\mathbb{P}^i_j\}^{i\in[N]}$ for all $j \in [M]$, such that any protocol among N agents needs to exchange $\Omega\left(\boxed{M N^2} \log(1/\varepsilon)\right)$ bits to achieve $\langle M, N, \varepsilon, \delta\rangle$-agreement on $\{f_j\}_{j\in[M]}$, for ε bounded below by $\min_{j\in[M]} \varepsilon_j$.*

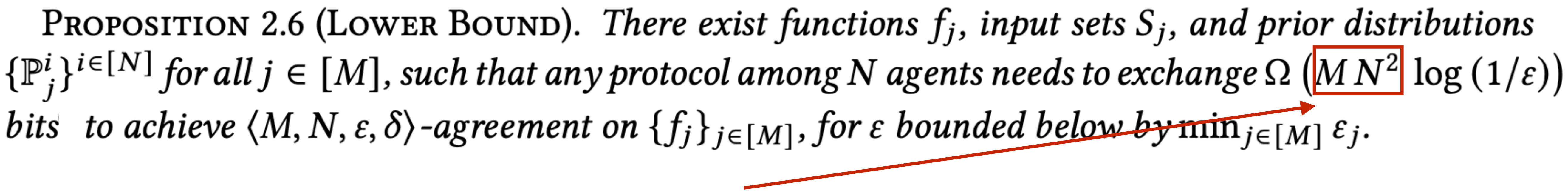
If we have a large number of tasks (*M*) or agents (*N*), then it is *impossible* to *always* align them efficiently, even if the agents are computationally unbounded.

We need to choose our tasks & agents wisely!



Implication: Brain-Computer Interfaces *won't* unilaterally solve the alignment problem because the *minimum* number of bits exchanged could be too large!

Fundamental Lower Bound: Unbounded Agent Setting



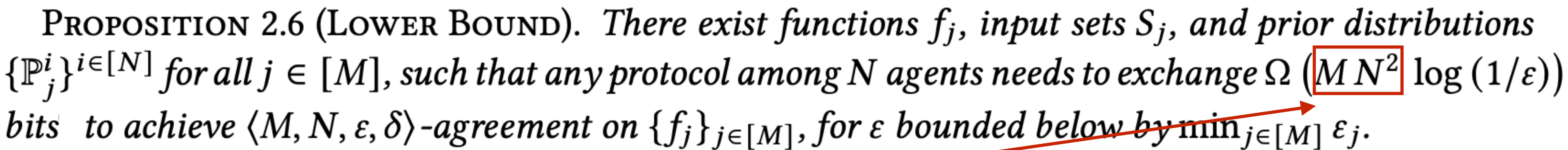
Proposition 2.6 (Lower Bound). *There exist functions f_j, input sets S_j, and prior distributions $\{\mathbb{P}^i_j\}^{i\in[N]}$ for all $j \in [M]$, such that any protocol among N agents needs to exchange $\Omega\left(\boxed{M N^2} \log(1/\varepsilon)\right)$ bits to achieve $\langle M, N, \varepsilon, \delta\rangle$-agreement on $\{f_j\}_{j\in[M]}$, for ε bounded below by $\min_{j\in[M]} \varepsilon_j$.*

If we have a large number of tasks (*M*) or agents (*N*), then it is *impossible* to *always* align them efficiently, even if the agents are computationally unbounded.

We need to choose our tasks & agents wisely!

Implication: Brain-Computer Interfaces *won't* unilaterally solve the alignment problem because the *minimum* number of bits exchanged could be too large!

Fundamental Lower Bound: Unbounded Agent Setting



Proposition 2.6 (Lower Bound). *There exist functions f_j, input sets S_j, and prior distributions $\{\mathbb{P}_j^i\}^{i\in[N]}$ for all $j \in [M]$, such that any protocol among N agents needs to exchange $\Omega\left(\boxed{M N^2} \log(1/\varepsilon)\right)$ bits to achieve $\langle M, N, \varepsilon, \delta\rangle$-agreement on $\{f_j\}_{j\in[M]}$, for ε bounded below by $\min_{j\in[M]} \varepsilon_j$.*

If we have a large number of tasks (*M*) or agents (*N*), then it is *impossible* to *always* align them efficiently, even if the agents are computationally unbounded.

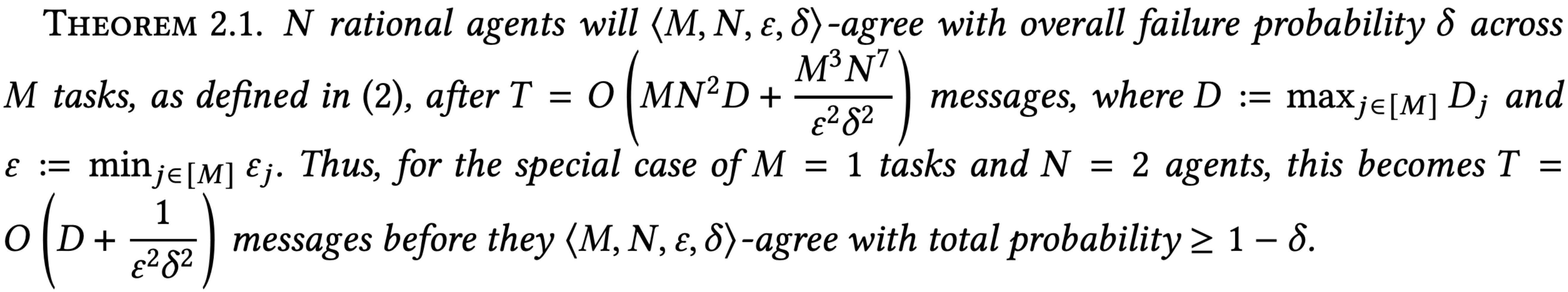
We need to choose our tasks & agents wisely!

Open Question (where NeuroAI can help!): What agent utility functions lead to incentives better for *us*?

Implication: Brain-Computer Interfaces *won't* unilaterally solve the alignment problem because the *minimum* number of bits exchanged could be too large!

Upper Bounds: Unbounded Agent Setting

Upper Bounds: Unbounded Agent Setting



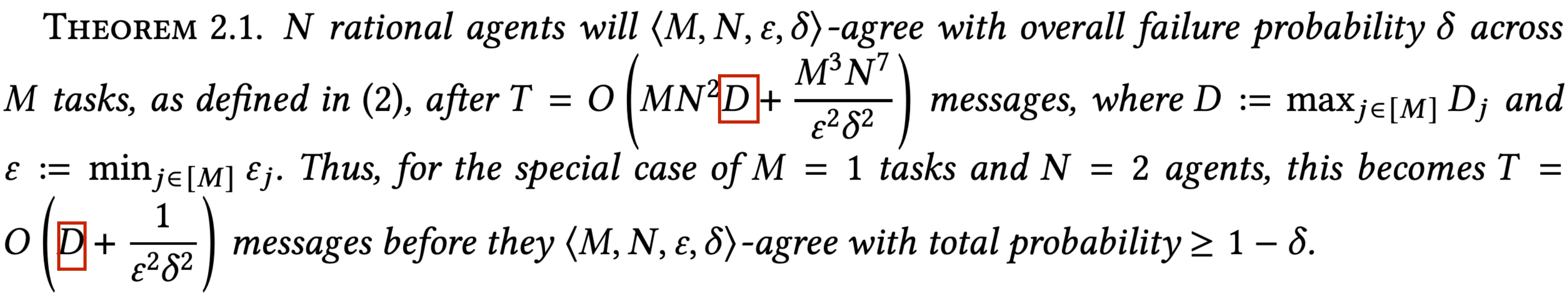
THEOREM 2.1. *N rational agents will $\langle M, N, \varepsilon, \delta \rangle$-agree with overall failure probability δ across M tasks, as defined in (2), after $T = O\left(MN^2D + \frac{M^3N^7}{\varepsilon^2\delta^2}\right)$ messages, where $D := \max_{j\in[M]} D_j$ and $\varepsilon := \min_{j\in[M]} \varepsilon_j$. Thus, for the special case of $M = 1$ tasks and $N = 2$ agents, this becomes $T = O\left(D + \frac{1}{\varepsilon^2\delta^2}\right)$ messages before they $\langle M, N, \varepsilon, \delta \rangle$-agree with total probability $\geq 1 - \delta$.*

Upper Bounds: Unbounded Agent Setting

Theorem 2.1. *N rational agents will $\langle M, N, \varepsilon, \delta\rangle$-agree with overall failure probability δ across M tasks, as defined in (2), after $T = O\left(MN^2\boxed{D} + \frac{M^3N^7}{\varepsilon^2\delta^2}\right)$ messages, where $D := \max_{j\in[M]} D_j$ and $\varepsilon := \min_{j\in[M]} \varepsilon_j$. Thus, for the special case of $M = 1$ tasks and $N = 2$ agents, this becomes $T = O\left(\boxed{D} + \frac{1}{\varepsilon^2\delta^2}\right)$ messages before they $\langle M, N, \varepsilon, \delta\rangle$-agree with total probability $\geq 1 - \delta$.*

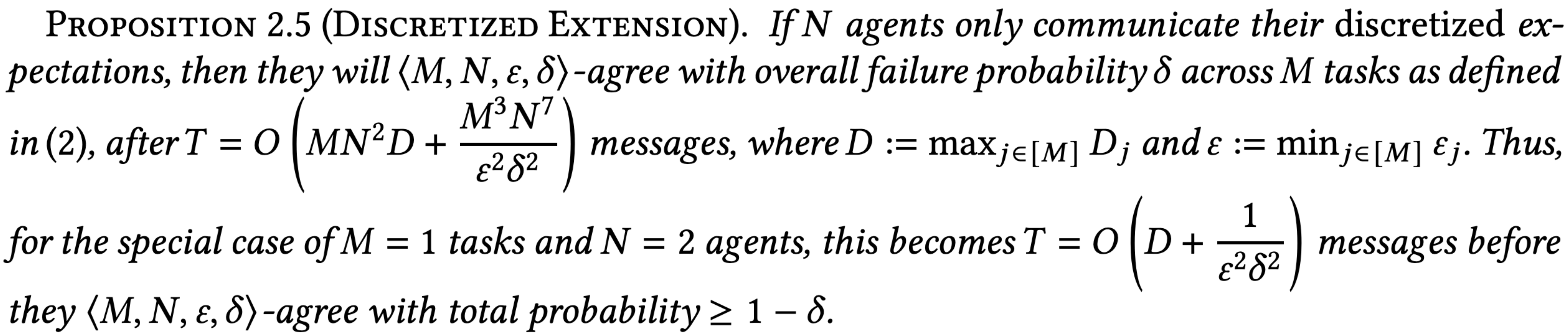
Linear in task state space size *D* (which is usually exponentially large in practice!)

Upper Bounds: Unbounded Agent Setting



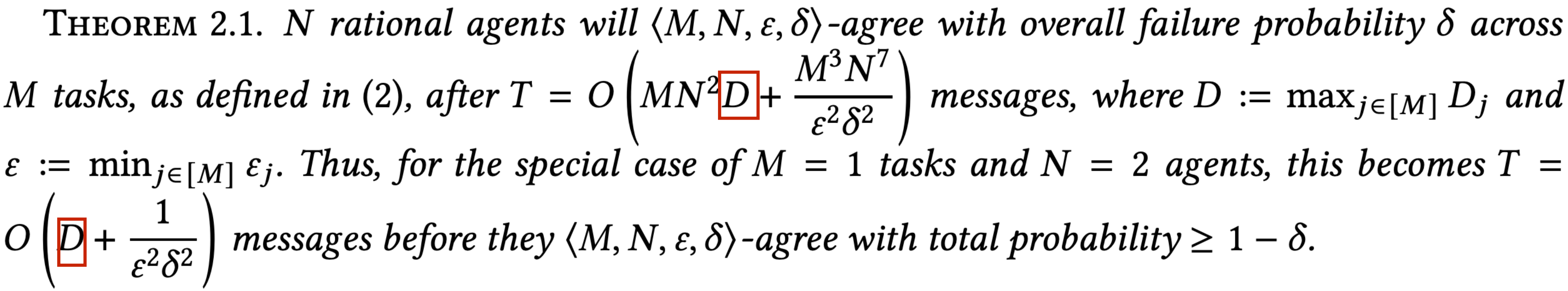
Theorem 2.1. *N rational agents will $\langle M, N, \varepsilon, \delta\rangle$-agree with overall failure probability δ across M tasks, as defined in* (2)*, after $T = O\left(MN^2\boxed{D} + \frac{M^3N^7}{\varepsilon^2\delta^2}\right)$ messages, where $D := \max_{j\in[M]} D_j$ and $\varepsilon := \min_{j\in[M]} \varepsilon_j$. Thus, for the special case of $M = 1$ tasks and $N = 2$ agents, this becomes $T = O\left(\boxed{D} + \frac{1}{\varepsilon^2\delta^2}\right)$ messages before they $\langle M, N, \varepsilon, \delta\rangle$-agree with total probability $\geq 1 - \delta$.*

Linear in task state space size *D* (which is usually exponentially large in practice!)



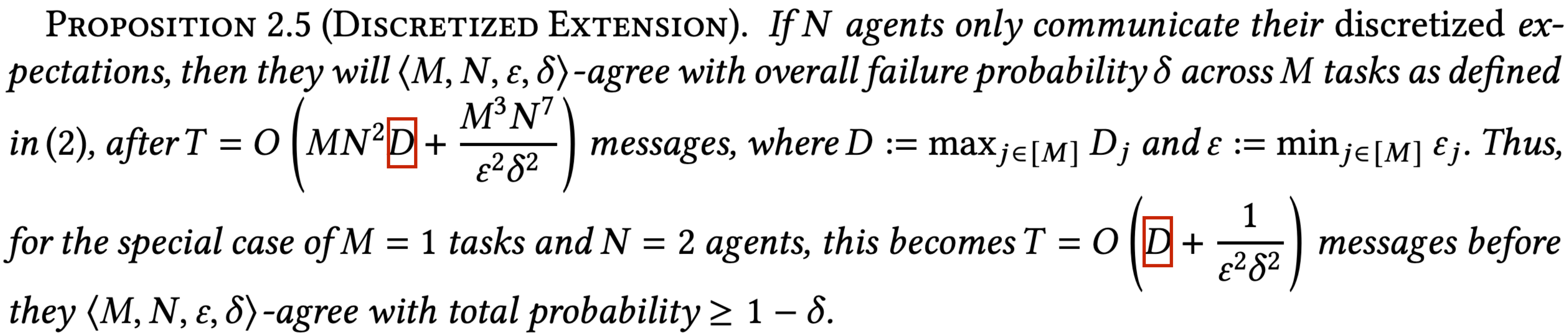
Proposition 2.5 (Discretized Extension). *If N agents only communicate their* discretized *expectations, then they will $\langle M, N, \varepsilon, \delta\rangle$-agree with overall failure probability δ across M tasks as defined in* (2)*, after $T = O\left(MN^2D + \frac{M^3N^7}{\varepsilon^2\delta^2}\right)$ messages, where $D := \max_{j\in[M]} D_j$ and $\varepsilon := \min_{j\in[M]} \varepsilon_j$. Thus, for the special case of $M = 1$ tasks and $N = 2$ agents, this becomes $T = O\left(D + \frac{1}{\varepsilon^2\delta^2}\right)$ messages before they $\langle M, N, \varepsilon, \delta\rangle$-agree with total probability $\geq 1 - \delta$.*

Upper Bounds: Unbounded Agent Setting



THEOREM 2.1. *N rational agents will $\langle M, N, \varepsilon, \delta \rangle$-agree with overall failure probability δ across M tasks, as defined in (2), after $T = O\left(MN^2\boxed{D} + \frac{M^3N^7}{\varepsilon^2\delta^2}\right)$ messages, where $D := \max_{j \in [M]} D_j$ and $\varepsilon := \min_{j \in [M]} \varepsilon_j$. Thus, for the special case of $M = 1$ tasks and $N = 2$ agents, this becomes $T = O\left(\boxed{D} + \frac{1}{\varepsilon^2\delta^2}\right)$ messages before they $\langle M, N, \varepsilon, \delta \rangle$-agree with total probability $\geq 1 - \delta$.*

Linear in task state space size *D* (which is usually exponentially large in practice!)



PROPOSITION 2.5 (DISCRETIZED EXTENSION). *If N agents only communicate their* discretized *expectations, then they will $\langle M, N, \varepsilon, \delta \rangle$-agree with overall failure probability δ across M tasks as defined in (2), after $T = O\left(MN^2\boxed{D} + \frac{M^3N^7}{\varepsilon^2\delta^2}\right)$ messages, where $D := \max_{j \in [M]} D_j$ and $\varepsilon := \min_{j \in [M]} \varepsilon_j$. Thus, for the special case of $M = 1$ tasks and $N = 2$ agents, this becomes $T = O\left(\boxed{D} + \frac{1}{\varepsilon^2\delta^2}\right)$ messages before they $\langle M, N, \varepsilon, \delta \rangle$-agree with total probability $\geq 1 - \delta$.*

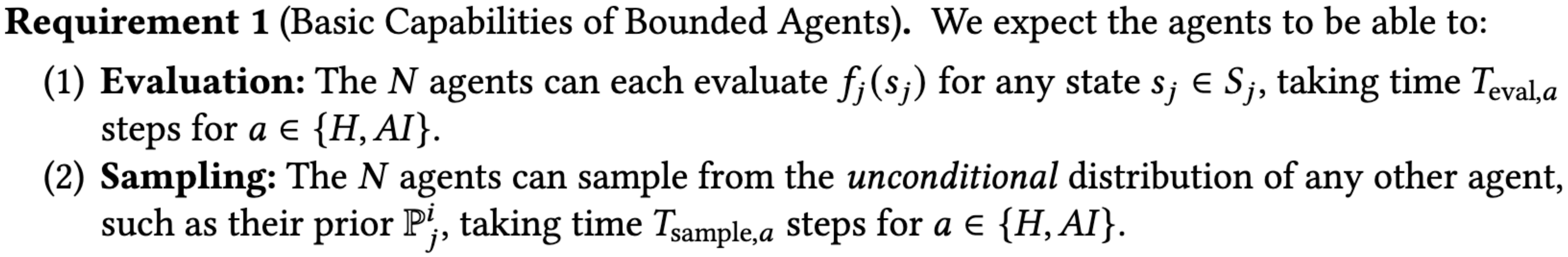
Discretized messages don't speed things up over real-valued messages

Bounded Agent Setting

What happens if the agents are computationally *bounded*, so messages no longer take *O(1)* time, and have noise in them (obfuscated intent)?

Bounded Agent Setting

What happens if the agents are computationally *bounded*, so messages no longer take *O(1)* time, and have noise in them (obfuscated intent)?

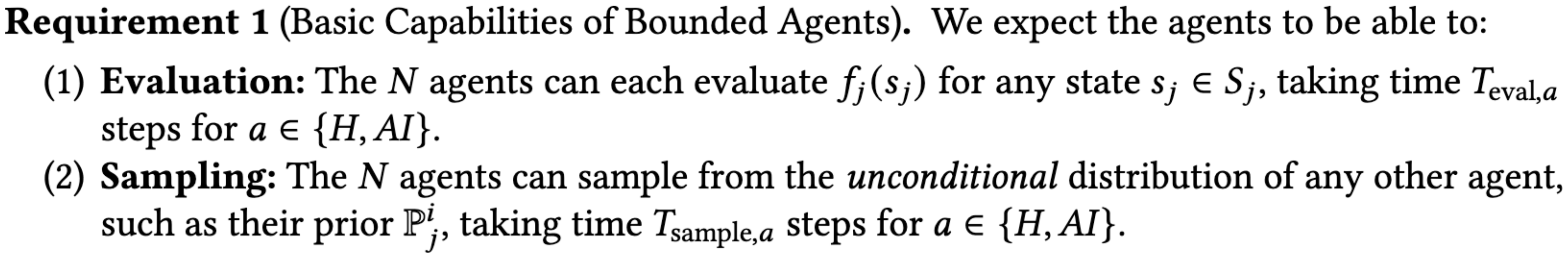


Requirement 1 (Basic Capabilities of Bounded Agents). We expect the agents to be able to:

1. **Evaluation:** The N agents can each evaluate $f_j(s_j)$ for any state $s_j \in S_j$, taking time $T_{\text{eval},a}$ steps for $a \in \{H, AI\}$.
2. **Sampling:** The N agents can sample from the *unconditional* distribution of any other agent, such as their prior $\mathbb{P}^i_j$, taking time $T_{\text{sample},a}$ steps for $a \in \{H, AI\}$.

Bounded Agent Setting

What happens if the agents are computationally *bounded*, so messages no longer take *O(1)* time, and have noise in them (obfuscated intent)?



Requirement 1 (Basic Capabilities of Bounded Agents). We expect the agents to be able to:

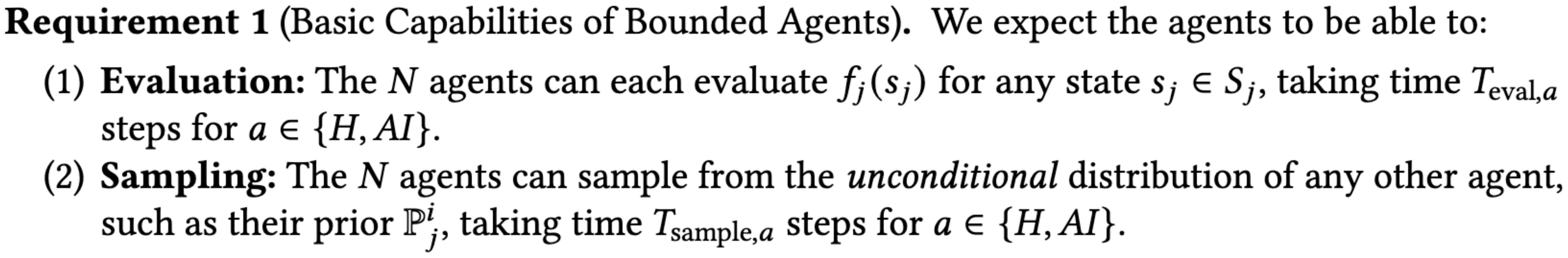
(1) **Evaluation:** The N agents can each evaluate $f_j(s_j)$ for any state $s_j \in S_j$, taking time $T_{\text{eval},a}$ steps for $a \in \{H, AI\}$.

(2) **Sampling:** The N agents can sample from the *unconditional* distribution of any other agent, such as their prior $\mathbb{P}^i_j$, taking time $T_{\text{sample},a}$ steps for $a \in \{H, AI\}$.

Intended to capture how querying a human is often more costly (in terms of time) than querying AI

Bounded Agent Setting

What happens if the agents are computationally *bounded*, so messages no longer take *O(1)* time, and have noise in them (obfuscated intent)?



Requirement 1 (Basic Capabilities of Bounded Agents). We expect the agents to be able to:

(1) **Evaluation:** The N agents can each evaluate $f_j(s_j)$ for any state $s_j \in S_j$, taking time $T_{\text{eval},a}$ steps for $a \in \{H, AI\}$.

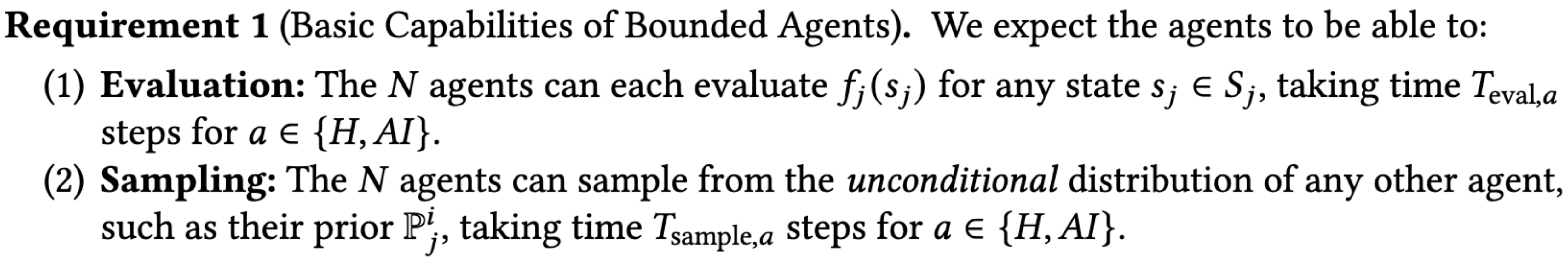
(2) **Sampling:** The N agents can sample from the *unconditional* distribution of any other agent, such as their prior $\mathbb{P}^i_j$, taking time $T_{\text{sample},a}$ steps for $a \in \{H, AI\}$.

Intended to capture how querying a human is often more costly (in terms of time) than querying AI

Note: Eval and sampling are black-boxes—agents learn through subroutines, not explicit descriptions. This reflects how we often recognize task completion without predefining execution steps.

Bounded Agent Setting

What happens if the agents are computationally *bounded*, so messages no longer take *O(1)* time, and have noise in them (obfuscated intent)?



Requirement 1 (Basic Capabilities of Bounded Agents). We expect the agents to be able to:

(1) **Evaluation:** The N agents can each evaluate $f_j(s_j)$ for any state $s_j \in S_j$, taking time $T_{\text{eval},a}$ steps for $a \in \{H, AI\}$.

(2) **Sampling:** The N agents can sample from the *unconditional* distribution of any other agent, such as their prior $\mathbb{P}^i_j$, taking time $T_{\text{sample},a}$ steps for $a \in \{H, AI\}$.

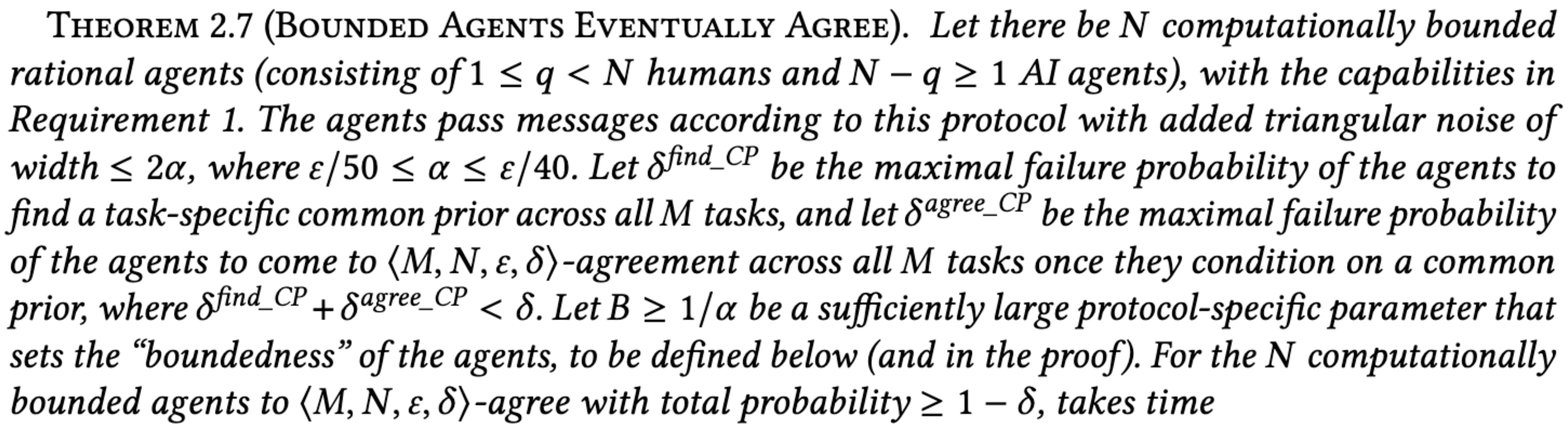
Intended to capture how querying a human is often more costly (in terms of time) than querying AI

<u>Note:</u> Eval and sampling are black-boxes—agents learn through subroutines, not explicit descriptions. This reflects how we often recognize task completion without predefining execution steps.

TL;DR: Exponential slowdown in task state space size (D)

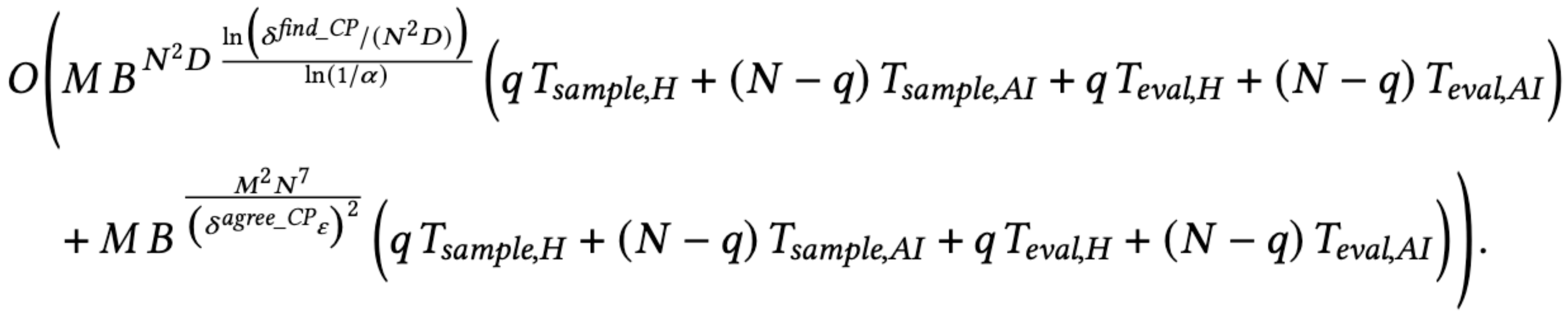
Main Results: Bounded Agent Setting

Main Results: Bounded Agent Setting



Theorem 2.7 (Bounded Agents Eventually Agree). *Let there be N computationally bounded rational agents (consisting of $1 \leq q < N$ humans and $N - q \geq 1$ AI agents), with the capabilities in Requirement 1. The agents pass messages according to this protocol with added triangular noise of width $\leq 2\alpha$, where $\varepsilon/50 \leq \alpha \leq \varepsilon/40$. Let δ^{find_CP} be the maximal failure probability of the agents to find a task-specific common prior across all M tasks, and let δ^{agree_CP} be the maximal failure probability of the agents to come to $\langle M, N, \varepsilon, \delta \rangle$-agreement across all M tasks once they condition on a common prior, where $\delta^{find\_CP} + \delta^{agree\_CP} < \delta$. Let $B \geq 1/\alpha$ be a sufficiently large protocol-specific parameter that sets the "boundedness" of the agents, to be defined below (and in the proof). For the N computationally bounded agents to $\langle M, N, \varepsilon, \delta \rangle$-agree with total probability $\geq 1 - \delta$, takes time*

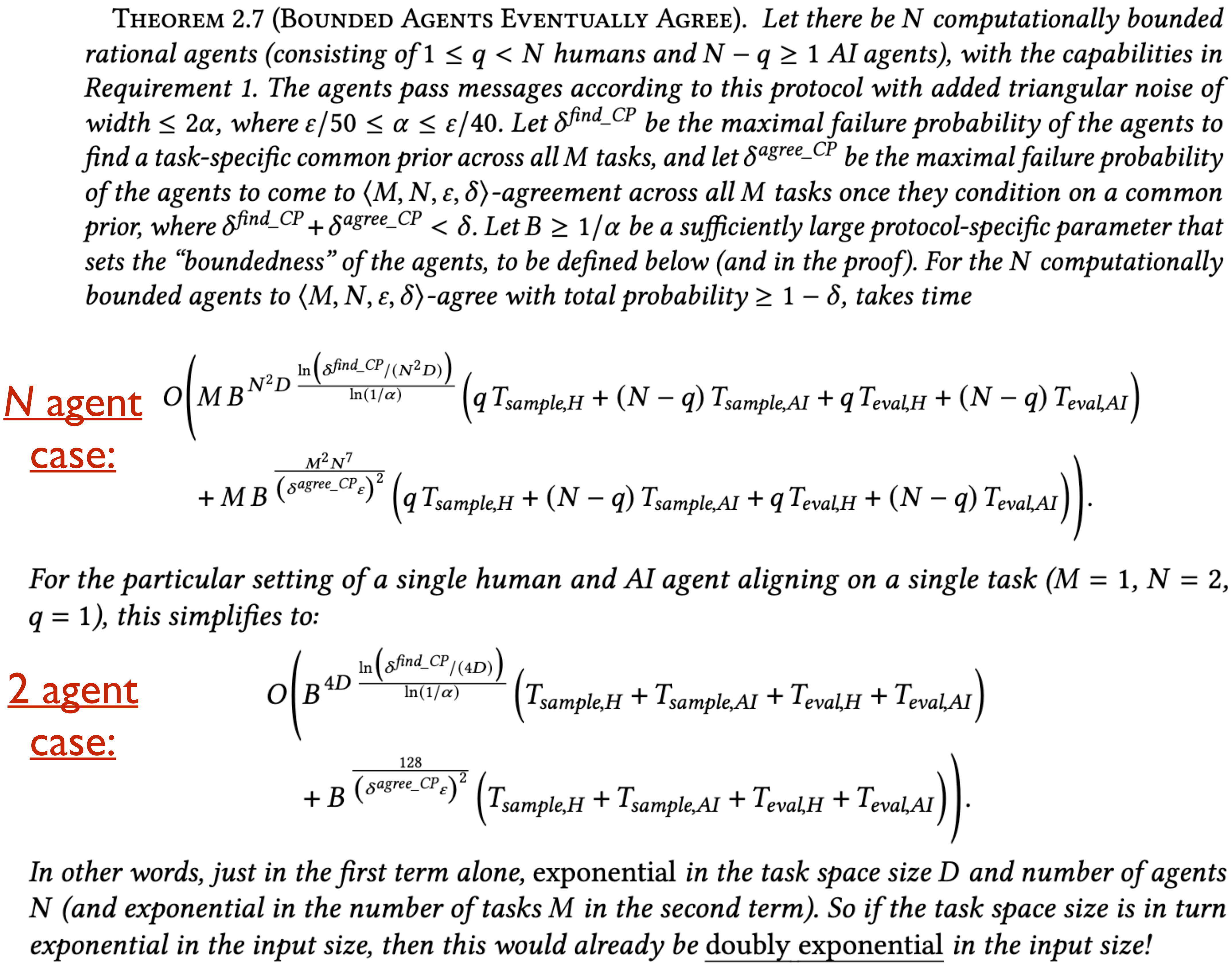
N agent case:



$$O\Bigg(M\,B^{N^2 D\,\frac{\ln\left(\delta^{find\_CP}/(N^2 D)\right)}{\ln(1/\alpha)}}\left(q\,T_{sample,H} + (N-q)\,T_{sample,AI} + q\,T_{eval,H} + (N-q)\,T_{eval,AI}\right)$$

$$+\,M\,B^{\frac{M^2 N^7}{\left(\delta^{agree\_CP}\varepsilon\right)^2}}\left(q\,T_{sample,H} + (N-q)\,T_{sample,AI} + q\,T_{eval,H} + (N-q)\,T_{eval,AI}\right)\Bigg).$$

Main Results: Bounded Agent Setting



THEOREM 2.7 (BOUNDED AGENTS EVENTUALLY AGREE). *Let there be N computationally bounded rational agents (consisting of $1 \leq q < N$ humans and $N - q \geq 1$ AI agents), with the capabilities in Requirement 1. The agents pass messages according to this protocol with added triangular noise of width $\leq 2\alpha$, where $\varepsilon/50 \leq \alpha \leq \varepsilon/40$. Let δ^{find_CP} be the maximal failure probability of the agents to find a task-specific common prior across all M tasks, and let δ^{agree_CP} be the maximal failure probability of the agents to come to $\langle M, N, \varepsilon, \delta \rangle$-agreement across all M tasks once they condition on a common prior, where $\delta^{find\_CP} + \delta^{agree\_CP} < \delta$. Let $B \geq 1/\alpha$ be a sufficiently large protocol-specific parameter that sets the "boundedness" of the agents, to be defined below (and in the proof). For the N computationally bounded agents to $\langle M, N, \varepsilon, \delta \rangle$-agree with total probability $\geq 1 - \delta$, takes time*

N agent case:

$$O\Bigg(M\,B^{N^2 D\,\frac{\ln\left(\delta^{find\_CP}/(N^2 D)\right)}{\ln(1/\alpha)}}\left(q\,T_{sample,H} + (N-q)\,T_{sample,AI} + q\,T_{eval,H} + (N-q)\,T_{eval,AI}\right)$$

$$+ M\,B^{\frac{M^2 N^7}{\left(\delta^{agree\_CP}\varepsilon\right)^2}}\left(q\,T_{sample,H} + (N-q)\,T_{sample,AI} + q\,T_{eval,H} + (N-q)\,T_{eval,AI}\right)\Bigg).$$

For the particular setting of a single human and AI agent aligning on a single task ($M = 1$, $N = 2$, $q = 1$), this simplifies to:

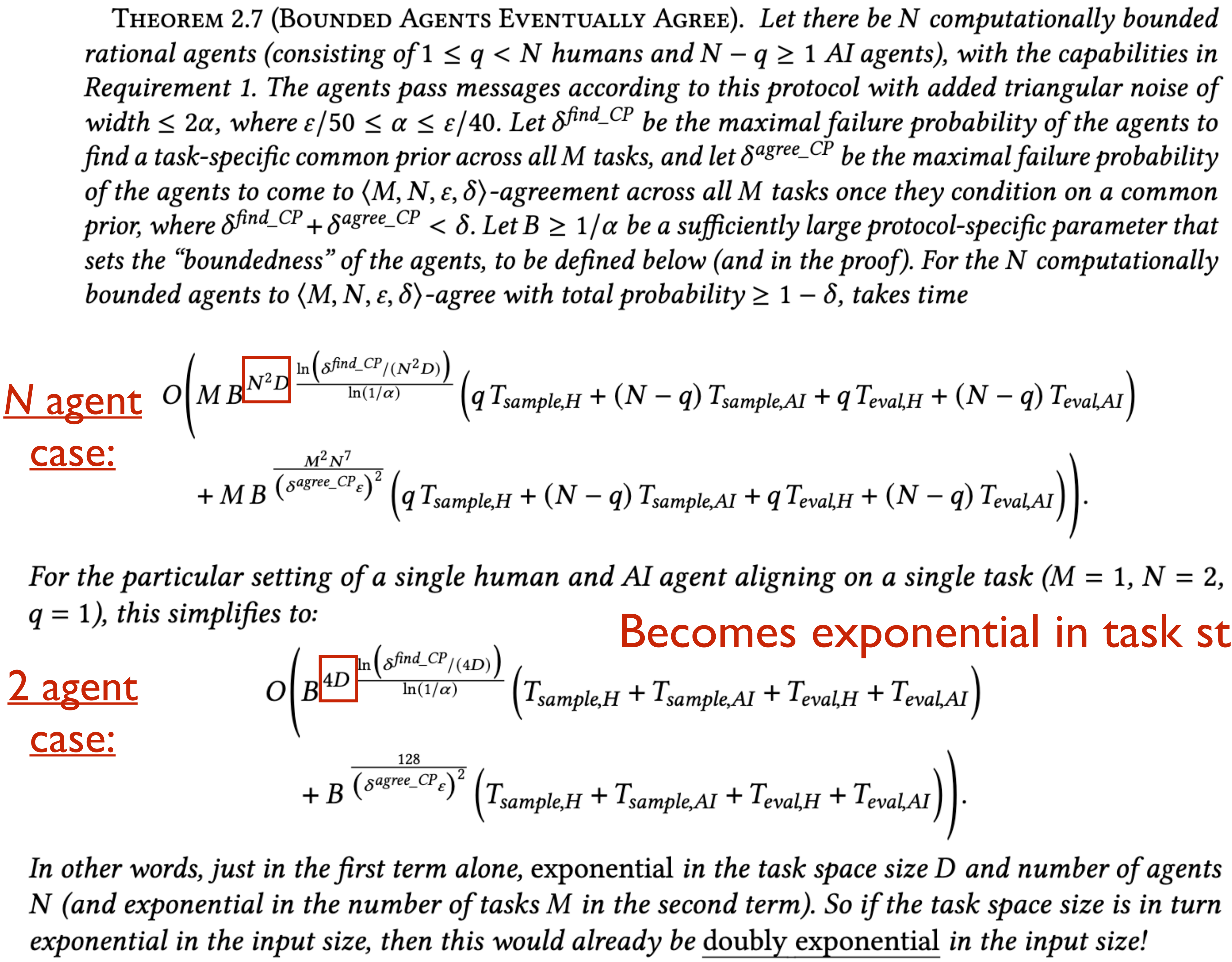
2 agent case:

$$O\Bigg(B^{4D\,\frac{\ln\left(\delta^{find\_CP}/(4D)\right)}{\ln(1/\alpha)}}\left(T_{sample,H} + T_{sample,AI} + T_{eval,H} + T_{eval,AI}\right)$$

$$+ B^{\frac{128}{\left(\delta^{agree\_CP}\varepsilon\right)^2}}\left(T_{sample,H} + T_{sample,AI} + T_{eval,H} + T_{eval,AI}\right)\Bigg).$$

In other words, just in the first term alone, exponential *in the task space size D and number of agents N (and exponential in the number of tasks M in the second term). So if the task space size is in turn exponential in the input size, then this would already be* doubly exponential *in the input size!*

Main Results: Bounded Agent Setting



THEOREM 2.7 (BOUNDED AGENTS EVENTUALLY AGREE). *Let there be N computationally bounded rational agents (consisting of $1 \le q < N$ humans and $N - q \ge 1$ AI agents), with the capabilities in Requirement 1. The agents pass messages according to this protocol with added triangular noise of width $\le 2\alpha$, where $\varepsilon/50 \le \alpha \le \varepsilon/40$. Let δ^{find_CP} be the maximal failure probability of the agents to find a task-specific common prior across all M tasks, and let δ^{agree_CP} be the maximal failure probability of the agents to come to $\langle M, N, \varepsilon, \delta\rangle$-agreement across all M tasks once they condition on a common prior, where $\delta^{find\_CP} + \delta^{agree\_CP} < \delta$. Let $B \ge 1/\alpha$ be a sufficiently large protocol-specific parameter that sets the "boundedness" of the agents, to be defined below (and in the proof). For the N computationally bounded agents to $\langle M, N, \varepsilon, \delta\rangle$-agree with total probability $\ge 1 - \delta$, takes time*

N agent case:

$$O\left(M\,B^{\boxed{N^2 D}\frac{\ln\left(\delta^{find\_CP}/(N^2 D)\right)}{\ln(1/\alpha)}}\left(q\,T_{sample,H} + (N-q)\,T_{sample,AI} + q\,T_{eval,H} + (N-q)\,T_{eval,AI}\right)\right.$$
$$\left. + M\,B^{\frac{M^2 N^7}{\left(\delta^{agree\_CP}\varepsilon\right)^2}}\left(q\,T_{sample,H} + (N-q)\,T_{sample,AI} + q\,T_{eval,H} + (N-q)\,T_{eval,AI}\right)\right).$$

For the particular setting of a single human and AI agent aligning on a single task ($M = 1$, $N = 2$, $q = 1$), this simplifies to:

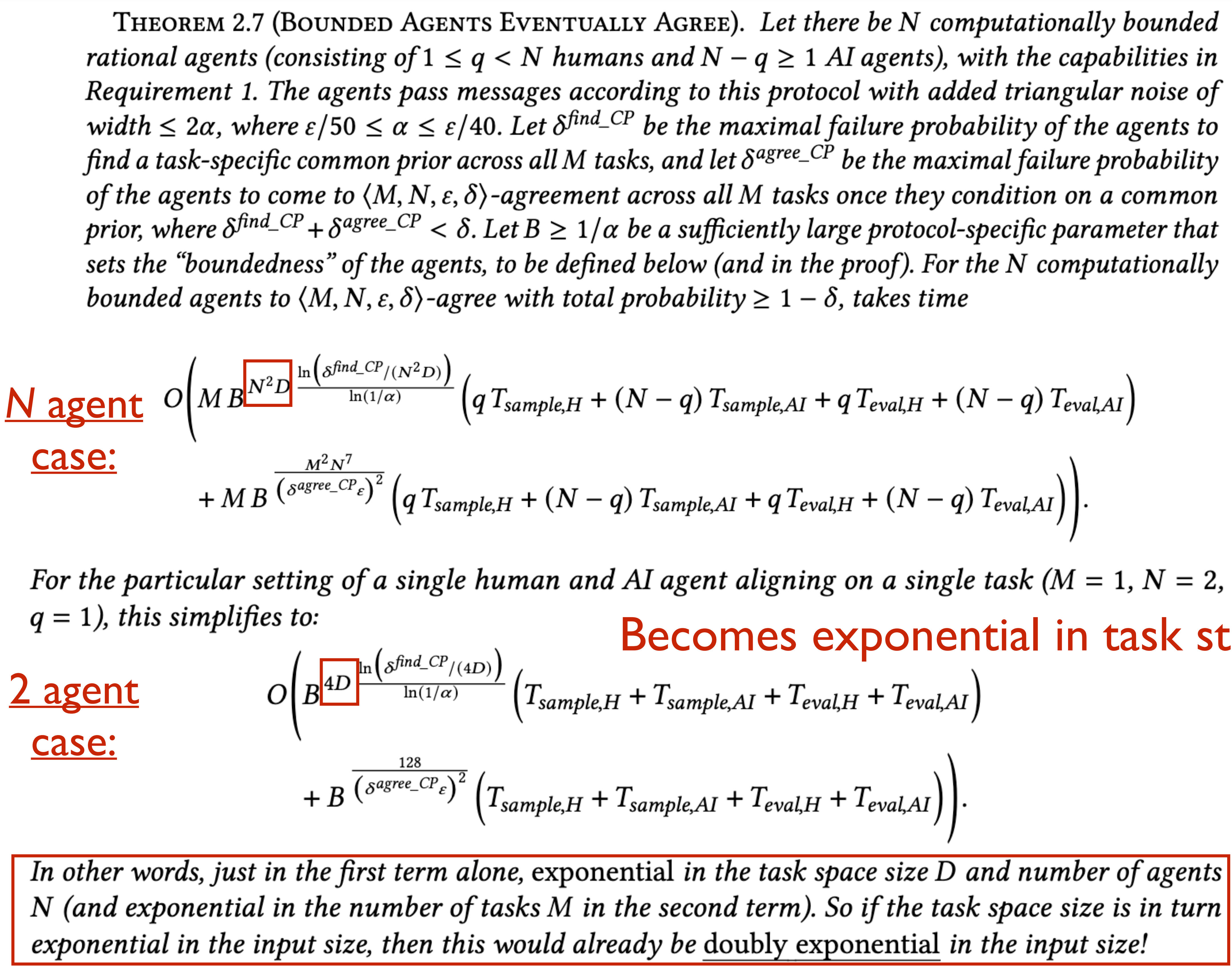
Becomes exponential in task state space *D*!

2 agent case:

$$O\left(B^{\boxed{4D}\frac{\ln\left(\delta^{find\_CP}/(4D)\right)}{\ln(1/\alpha)}}\left(T_{sample,H} + T_{sample,AI} + T_{eval,H} + T_{eval,AI}\right)\right.$$
$$\left. + B^{\frac{128}{\left(\delta^{agree\_CP}\varepsilon\right)^2}}\left(T_{sample,H} + T_{sample,AI} + T_{eval,H} + T_{eval,AI}\right)\right).$$

In other words, just in the first term alone, exponential *in the task space size D and number of agents N (and exponential in the number of tasks M in the second term). So if the task space size is in turn exponential in the input size, then this would already be* doubly exponential *in the input size!*

Main Results: Bounded Agent Setting



Theorem 2.7 (Bounded Agents Eventually Agree). *Let there be N computationally bounded rational agents (consisting of $1 \le q < N$ humans and $N - q \ge 1$ AI agents), with the capabilities in Requirement 1. The agents pass messages according to this protocol with added triangular noise of width $\le 2\alpha$, where $\varepsilon/50 \le \alpha \le \varepsilon/40$. Let δ^{find_CP} be the maximal failure probability of the agents to find a task-specific common prior across all M tasks, and let δ^{agree_CP} be the maximal failure probability of the agents to come to $\langle M, N, \varepsilon, \delta \rangle$-agreement across all M tasks once they condition on a common prior, where $\delta^{find\_CP} + \delta^{agree\_CP} < \delta$. Let $B \ge 1/\alpha$ be a sufficiently large protocol-specific parameter that sets the "boundedness" of the agents, to be defined below (and in the proof). For the N computationally bounded agents to $\langle M, N, \varepsilon, \delta \rangle$-agree with total probability $\ge 1 - \delta$, takes time*

N agent case:

$$O\left(M\,B^{\boxed{N^2D}\frac{\ln\left(\delta^{find\_CP}/(N^2D)\right)}{\ln(1/\alpha)}}\left(q\,T_{sample,H} + (N-q)\,T_{sample,AI} + q\,T_{eval,H} + (N-q)\,T_{eval,AI}\right)\right.$$

$$\left. + \, M\,B^{\frac{M^2N^7}{\left(\delta^{agree\_CP}\varepsilon\right)^2}}\left(q\,T_{sample,H} + (N-q)\,T_{sample,AI} + q\,T_{eval,H} + (N-q)\,T_{eval,AI}\right)\right).$$

For the particular setting of a single human and AI agent aligning on a single task ($M = 1$, $N = 2$, $q = 1$), this simplifies to:

Becomes exponential in task state space *D*!

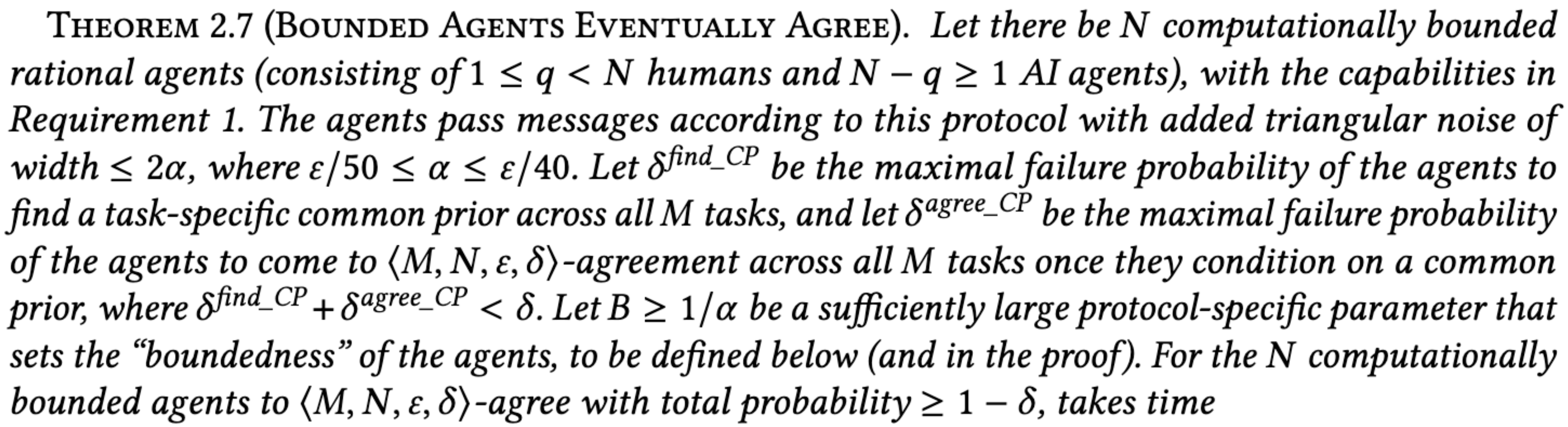
2 agent case:

$$O\left(B^{\boxed{4D}\frac{\ln\left(\delta^{find\_CP}/(4D)\right)}{\ln(1/\alpha)}}\left(T_{sample,H} + T_{sample,AI} + T_{eval,H} + T_{eval,AI}\right)\right.$$

$$\left. + \, B^{\frac{128}{\left(\delta^{agree\_CP}\varepsilon\right)^2}}\left(T_{sample,H} + T_{sample,AI} + T_{eval,H} + T_{eval,AI}\right)\right).$$

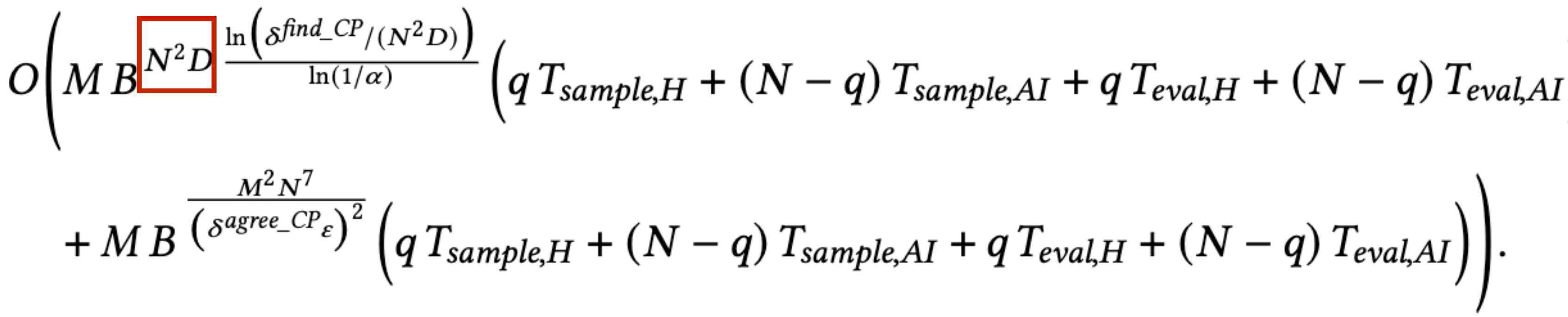
In other words, just in the first term alone, exponential *in the task space size D and number of agents N (and exponential in the number of tasks M in the second term). So if the task space size is in turn exponential in the input size, then this would already be* doubly exponential *in the input size!*

Main Results: Bounded Agent Setting



Theorem 2.7 (Bounded Agents Eventually Agree). *Let there be N computationally bounded rational agents (consisting of $1 \leq q < N$ humans and $N - q \geq 1$ AI agents), with the capabilities in Requirement 1. The agents pass messages according to this protocol with added triangular noise of width $\leq 2\alpha$, where $\varepsilon/50 \leq \alpha \leq \varepsilon/40$. Let δ^{find_CP} be the maximal failure probability of the agents to find a task-specific common prior across all M tasks, and let δ^{agree_CP} be the maximal failure probability of the agents to come to $\langle M, N, \varepsilon, \delta\rangle$-agreement across all M tasks once they condition on a common prior, where $\delta^{find\_CP} + \delta^{agree\_CP} < \delta$. Let $B \geq 1/\alpha$ be a sufficiently large protocol-specific parameter that sets the "boundedness" of the agents, to be defined below (and in the proof). For the N computationally bounded agents to $\langle M, N, \varepsilon, \delta\rangle$-agree with total probability $\geq 1 - \delta$, takes time*

N agent case:



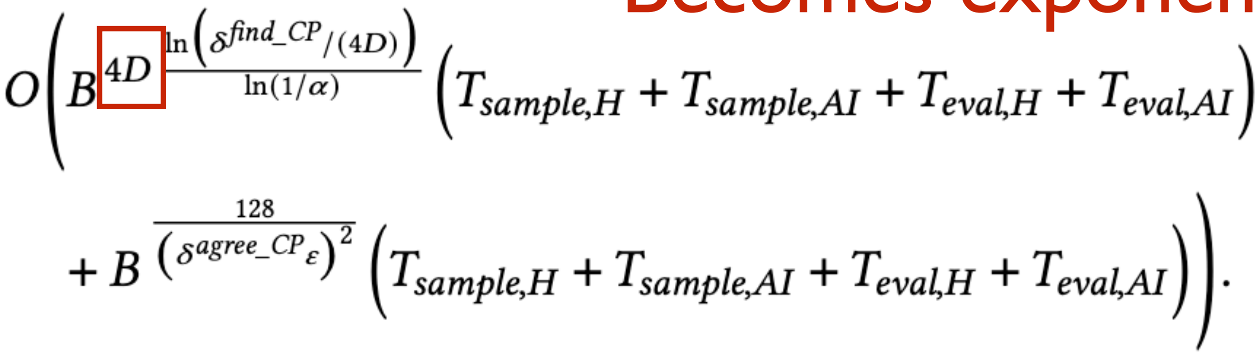
$$O\left(M\,B^{\boxed{N^2 D}\,\frac{\ln\left(\delta^{find\_CP}/(N^2 D)\right)}{\ln(1/\alpha)}}\left(q\,T_{sample,H} + (N-q)\,T_{sample,AI} + q\,T_{eval,H} + (N-q)\,T_{eval,AI}\right) + M\,B^{\frac{M^2 N^7}{\left(\delta^{agree\_CP}\varepsilon\right)^2}}\left(q\,T_{sample,H} + (N-q)\,T_{sample,AI} + q\,T_{eval,H} + (N-q)\,T_{eval,AI}\right)\right).$$

Obviously this is all bad, but humor me for a moment… just how bad can it get exactly?

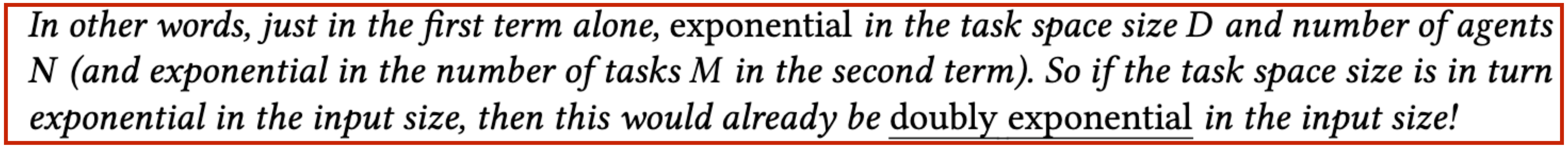
$q = 1$), this simplifies to:

Becomes exponential in task state space *D*!

2 agent case:



$$O\left(B^{\boxed{4D}\,\frac{\ln\left(\delta^{find\_CP}/(4D)\right)}{\ln(1/\alpha)}}\left(T_{sample,H} + T_{sample,AI} + T_{eval,H} + T_{eval,AI}\right) + B^{\frac{128}{\left(\delta^{agree\_CP}\varepsilon\right)^2}}\left(T_{sample,H} + T_{sample,AI} + T_{eval,H} + T_{eval,AI}\right)\right).$$



In other words, just in the first term alone, exponential *in the task space size D and number of agents N (and exponential in the number of tasks M in the second term). So if the task space size is in turn exponential in the input size, then this would already be doubly exponential in the input size!*

Total Bayesian Wannabe

Total Bayesian Wannabe

What if the bounded agents want to pass a "Bayesian Turing Test" of sorts: Namely, act indistinguishably from an unbounded Bayesian across all *M* tasks, as refereed by a watchful unbounded Bayesian?

Total Bayesian Wannabe

What if the bounded agents want to pass a "Bayesian Turing Test" of sorts: Namely, act indistinguishably from an unbounded Bayesian across all *M* tasks, as refereed by a watchful unbounded Bayesian?

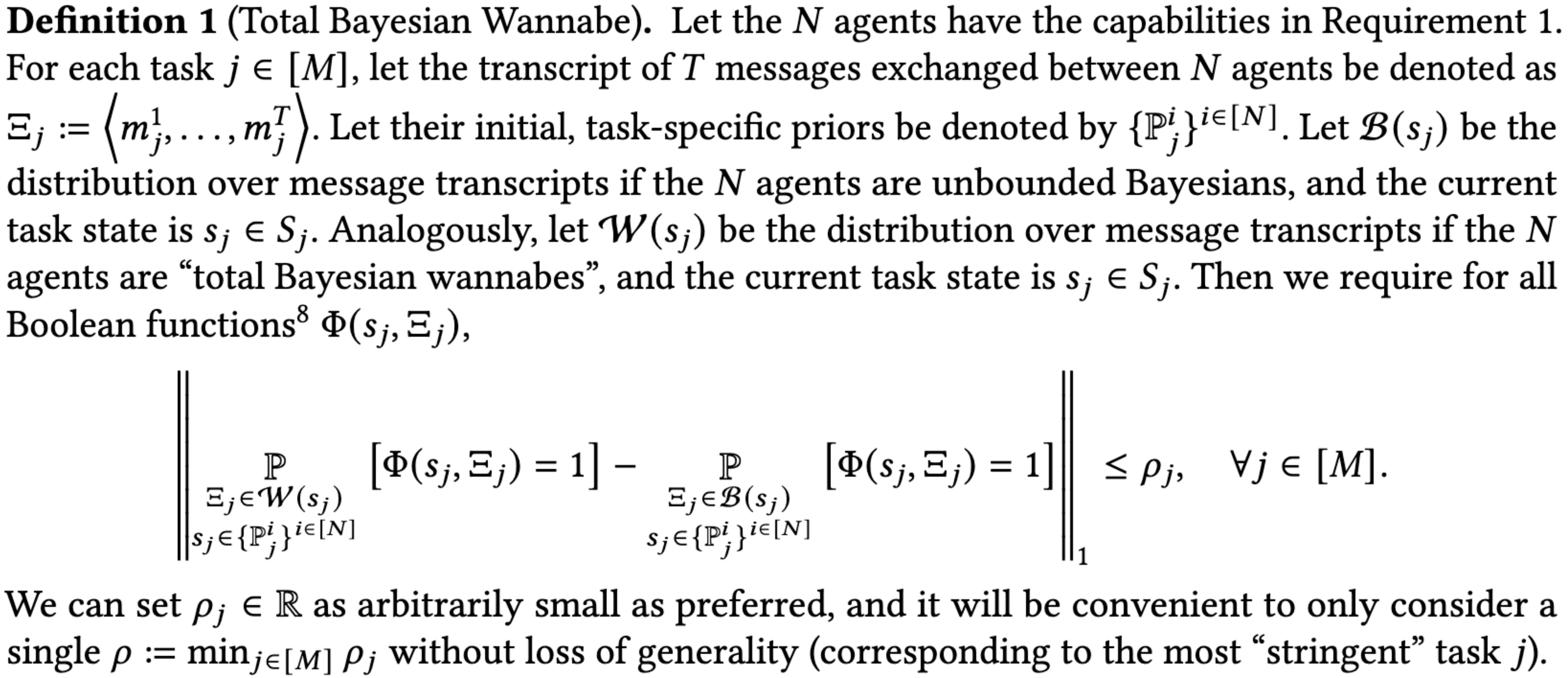
We will call them "Total Bayesian Wannabes"

Total Bayesian Wannabe

What if the bounded agents want to pass a "Bayesian Turing Test" of sorts: Namely, act indistinguishably from an unbounded Bayesian across all *M* tasks, as refereed by a watchful unbounded Bayesian?

We will call them "Total Bayesian Wannabes"

If interested, the technical definition is here:



Definition 1 (Total Bayesian Wannabe). Let the N agents have the capabilities in Requirement 1. For each task $j \in [M]$, let the transcript of T messages exchanged between N agents be denoted as $\Xi_j := \left\langle m_j^1, \ldots, m_j^T \right\rangle$. Let their initial, task-specific priors be denoted by $\{\mathbb{P}_j^i\}^{i\in[N]}$. Let $\mathcal{B}(s_j)$ be the distribution over message transcripts if the N agents are unbounded Bayesians, and the current task state is $s_j \in S_j$. Analogously, let $\mathcal{W}(s_j)$ be the distribution over message transcripts if the N agents are "total Bayesian wannabes", and the current task state is $s_j \in S_j$. Then we require for all Boolean functions[8] $\Phi(s_j, \Xi_j)$,

$$\left\| \underset{\substack{\Xi_j \in \mathcal{W}(s_j) \\ s_j \in \{\mathbb{P}_j^i\}^{i\in[N]}}}{\mathbb{P}} \left[\Phi(s_j, \Xi_j) = 1\right] - \underset{\substack{\Xi_j \in \mathcal{B}(s_j) \\ s_j \in \{\mathbb{P}_j^i\}^{i\in[N]}}}{\mathbb{P}} \left[\Phi(s_j, \Xi_j) = 1\right] \right\|_1 \leq \rho_j, \quad \forall j \in [M].$$

We can set $\rho_j \in \mathbb{R}$ as arbitrarily small as preferred, and it will be convenient to only consider a single $\rho := \min_{j\in[M]} \rho_j$ without loss of generality (corresponding to the most "stringent" task j).

Total Bayesian Wannabes Totally Wanna Agree If They Have Enough Time

$$O\left(\frac{(1100)^{\frac{2097152}{(1/4)^6}}}{(1/2)^{\frac{256}{(1/4)^2}}}\right) = O\left(1.31 \times 10^{26125365467}\right)$$

If the agents are computationally *bounded*, this can currently take more subroutine calls than the number of atoms in the observable universe! (~4.8 x 10^{79})

Takeaways

Takeaways

Even once we have capable agents, AI alignment *can* be computationally infeasible *in general.*

Takeaways

Even once we have capable agents, AI alignment *can* be computationally infeasible *in general.*

In fact, we showed that alignment is fundamentally constrained by 3 quantities: **the number of tasks (*M*)**, **agents (*N*)**, and **task state space size (*D*)**.

Takeaways

Even once we have capable agents, AI alignment *can* be computationally infeasible *in general.*

In fact, we showed that alignment is fundamentally constrained by 3 quantities: **the number of tasks (*M*)**, **agents (*N*)**, and **task state space size (*D*)**. The setting we consider here is a "best-case scenario" in some sense, and we already run into inefficiencies here.

Takeaways

Even once we have capable agents, AI alignment *can* be computationally infeasible *in general.*

In fact, we showed that alignment is fundamentally constrained by 3 quantities: **the number of tasks (*M*)**, **agents (*N*)**, and **task state space size (*D*)**. The setting we consider here is a "best-case scenario" in some sense, and we already run into inefficiencies here.

Thus, alignment might **<u>not</u>** be a scalable approach to AI safety in many settings, as it requires a lot of structure to be provably efficient.

Takeaways

Even once we have capable agents, AI alignment *can* be computationally infeasible *in general.*

In fact, we showed that alignment is fundamentally constrained by 3 quantities: **the number of tasks (*M*)**, **agents (*N*)**, and **task state space size (*D*)**. The setting we consider here is a "best-case scenario" in some sense, and we already run into inefficiencies here.

Thus, alignment might **<u>not</u>** be a scalable approach to AI safety in many settings, as it requires a lot of structure to be provably efficient. For example, we prescribe the following prescriptions for *specific* problems to avoid intractabilities:

Prescriptions for Practice

Prescriptions for Practice

1. Choose your tasks & agents wisely.

Prescriptions for Practice

1. Choose your tasks & agents wisely. Cut down on task space, e.g. don't always steer the base model directly, but "funnel" it through a smaller task space when possible.

Prescriptions for Practice

1. Choose your tasks & agents wisely. Cut down on task space, e.g. don't always steer the base model directly, but "funnel" it through a smaller task space when possible.

2. Leverage inductive biases of the task structure & agent architecture (encourages "efficient sampleability of posteriors")

Prescriptions for Practice

1. Choose your tasks & agents wisely. Cut down on task space, e.g. don't always steer the base model directly, but "funnel" it through a smaller task space when possible.

2. Leverage inductive biases of the task structure & agent architecture (encourages "efficient sampleability of posteriors")

3. Pretrain on human preferences (encourages a "common prior"): **Where NeuroAI can help!**

Prescriptions for Practice

1. Choose your tasks & agents wisely. Cut down on task space, e.g. don't always steer the base model directly, but "funnel" it through a smaller task space when possible.

2. Leverage inductive biases of the task structure & agent architecture (encourages "efficient sampleability of posteriors")

3. Pretrain on human preferences (encourages a "common prior"): **Where NeuroAI can help!**

4. Agents minimally should have 3 features: **bounded theory of mind**, **memory**, and **rationality**. Constitutes a "sufficiently safe" agent in this context to prove alignment guarantees.

Prescriptions for Practice

1. Choose your tasks & agents wisely. Cut down on task space, e.g. don't always steer the base model directly, but "funnel" it through a smaller task space when possible.

2. Leverage inductive biases of the task structure & agent architecture (encourages "efficient sampleability of posteriors")

3. Pretrain on human preferences (encourages a "common prior"): **Where NeuroAI can help!**

4. Agents minimally should have 3 features: **bounded theory of mind**, **memory**, and **rationality**. Constitutes a "sufficiently safe" agent in this context to prove alignment guarantees.

5. <u>Constraints on Obfuscated Intent</u>: Not all communication noise is an equally good choice to ensure alignment! (e.g. uniform noise won't work)

Prescriptions for Practice

1. Choose your tasks & agents wisely. Cut down on task space, e.g. don't always steer the base model directly, but "funnel" it through a smaller task space when possible.

2. Leverage inductive biases of the task structure & agent architecture (encourages "efficient sampleability of posteriors")

3. Pretrain on human preferences (encourages a "common prior"): **Where NeuroAI can help!**

4. Agents minimally should have 3 features: **bounded theory of mind**, **memory**, and **rationality**. Constitutes a "sufficiently safe" agent in this context to prove alignment guarantees.

5. Constraints on Obfuscated Intent: Not all communication noise is an equally good choice to ensure alignment! (e.g. uniform noise won't work)

Check out the paper for lots more details for each of these! (pp. 15-17)

Why Care?

We live in a society where people do not agree most of the time, but nothing "civilization-ending" has happened (yet)?

Why Care?

We live in a society where people do not agree most of the time, but nothing "civilization-ending" has happened (yet)?

Not *all* tasks require high agreement to avoid catastrophe: e.g. making a sandwich vs. running a nuclear power plant.

Why Care?

We live in a society where people do not agree most of the time, but nothing "civilization-ending" has happened (yet)?

Not *all* tasks require high agreement to avoid catastrophe: e.g. making a sandwich vs. running a nuclear power plant.

AI is something we are *intentionally* creating, so we should hold it to a higher standard than we do for other humans.

Why Care?

We live in a society where people do not agree most of the time, but nothing "civilization-ending" has happened (yet)?

Not *all* tasks require high agreement to avoid catastrophe: e.g. making a sandwich vs. running a nuclear power plant.

AI is something we are *intentionally* creating, so we should hold it to a higher standard than we do for other humans. Ensuring better incentives for AI agents that we can mostly agree on? (open question)

Why Care?

We live in a society where people do not agree most of the time, but nothing "civilization-ending" has happened (yet)?

Not *all* tasks require high agreement to avoid catastrophe: e.g. making a sandwich vs. running a nuclear power plant.

AI is something we are *intentionally* creating, so we should hold it to a higher standard than we do for other humans. Ensuring better incentives for AI agents that we can mostly agree on? (open question)

In some sense, the Pareto-optimal "worst case" is that if humans fail to sustain themselves for various reasons, at least we have something that carries on our intellectual legacy:

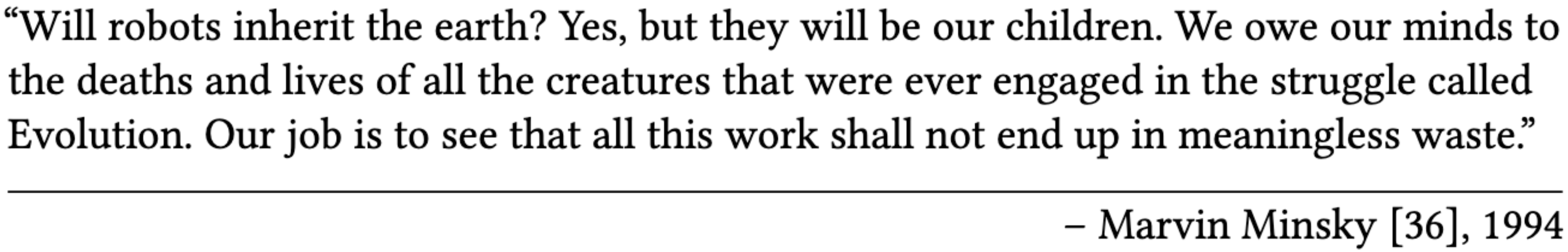
Why Care?

We live in a society where people do not agree most of the time, but nothing "civilization-ending" has happened (yet)?

Not *all* tasks require high agreement to avoid catastrophe: e.g. making a sandwich vs. running a nuclear power plant.

AI is something we are *intentionally* creating, so we should hold it to a higher standard than we do for other humans. Ensuring better incentives for AI agents that we can mostly agree on? (open question)

In some sense, the Pareto-optimal "worst case" is that if humans fail to sustain themselves for various reasons, at least we have something that carries on our intellectual legacy:



"Will robots inherit the earth? Yes, but they will be our children. We owe our minds to the deaths and lives of all the creatures that were ever engaged in the struggle called Evolution. Our job is to see that all this work shall not end up in meaningless waste."

– Marvin Minsky [36], 1994

Contact

Paper: https://arxiv.org/abs/2502.05934

Long Form Summary:

Contact:

✉ anayebi@cs.cmu.edu

𝕏 @aran_nayebi

🦋 @anayebi.bsky.social

🌐 https://cs.cmu.edu/~anayebi

Carnegie Mellon
SCHOOL OF COMPUTER SCIENCE

Funding:

Burroughs Wellcome Fund
CASI Award

Google Robotics Award