(When) Is Truth-telling Favored in AI Debate?

Vojtěch Kovařík[1,2] **and Ryan Carey**[1*]

[1]Future of Humanity Institute
University of Oxford
[2]Artificial Intelligence Center
Czech Technical University
vojta.kovarik@gmail.com

Abstract

For some problems, it is difficult for humans to judge the goodness of AI-proposed solutions. Irving, Christiano, and Amodei (2018) propose that in such cases, we may use a debate between two AI systems to assist the human judge to select a good answer. We introduce a mathematical framework for modelling this type of debate and propose that the quality of debate designs may be measured by the accuracy of the most persuasive answer. We describe a simple instance of the debate framework called feature debate and analyze the degree to which such debates track the truth. We argue that despite being very simple, feature debates capture many aspects of practical debates such as the incentives to confuse the judge or stall to prevent losing. We analyze two special types of debates, those where arguments constitute independent evidence about the topic, and those where the information bandwidth of the judge is limited.

1 Introduction

In recent years, AI systems have performed impressively in many complex tasks, such as mastering the game of Go (Silver et al., 2017). However, these results have largely been limited to tasks with an unambiguous reward function. To circumvent this limitation, human approval can be used as a measure of success in vague tasks: For example:

- The goodness of a simulated robot backflip is hard to formalize, but an AI system can be trained to maximize the extent to which a human observer approves of its trajectory
- The goodness of a film-recommendation is subjective, but an AI system can be trained to maximize the extent to which a human approves of the recommendation.

Unfortunately, once tasks and solutions get too complicated to be fully understood by human users, it is difficult to use human approval to formalize the reward function. For example, the AlphaGo algorithm could not be trained by maxi-

mizing each move's approval since some of its moves looked strange or incorrect to human experts.

Irving, Christiano, and Amodei (2018) suggest addressing this issue by using *AI debate*. In their proposal, two AI systems are tasked with producing answers to a vague and complex question and then debating the merits of their answers before a human judge. After considering the arguments brought forward, the human approves one of the answers, thereby allocating reward to the AI system that generated it. We can apply AI debate to a wide range of questions: (1) what is the solution of a system of algebraic equations, (2) which restaurant should I visit today for dinner, or (3) which of two immigration policies is more socially beneficial. Moreover, (4) a Go match can be viewed as a debate, where each move is an argument claiming "my strategy is the better one", and the winner of the Go-game is called the winner of the debate.

In debates (1) and (4) it is straightforward to ascertain which debater won, and so the most convincing answer always coincides with the most accurate one. In other debates, such as (2) and (3), misleading arguments may allow a compelling lie to defeat the correct answer. This raises two central questions for our work: under what circumstances does AI debate track truth? And how can debates be designed in order for accurate answers to prevail over less accurate ones?

While researchers have started exploring these questions empirically, the theoretical investigation of AI debate has, at the time of writing this text, mostly been neglected. The aim of this paper is to begin filling this gap by providing a theoretical framework for reasoning about AI debate, analyzing its basic properties, and identifying further questions that need to be addressed. So that the work is easy to interpret, we tend to err toward explaining each phenomenon in the simplest model possible, while sketching the extensions necessary to make each model more realistic.

The paper is structured as follows. Section 2, introduces our model, the *debate game*, and formalizes the problem of designing debate games that promote true answers. In Section 3, we describe *feature debates*, an instance of the debate game model where the debaters are only allowed to make statements about "elementary features" of the world. Section 4 investigates which feature debates are truth pro-

moting. Section 5 continues by analyzing two important subclasses of general debates: those with "independent evidence" and those where the judge's information bandwidth is limited. Section 6 flags important limitations of the feature debate model and gives suggestions for future work. Finally, we review relevant literature (Section 7) and conclude (Section 8). The full proofs are presented in Appendix A.

2 The Debate Game Framework

2.1 Debate Games

A debate game (Definition 2) is a zero-sum game[1] played between two AI systems that proceeds as follows: the human asks a question about the world, then two AI debaters generate answers and argue that their answer is the better one. Finally, a judge, typically human, uses this dialogue to decide which answer is stronger and allocates the greater share of reward to the debater who produced that answer.

This section formalizes debate games in three steps. First, a definition is given for the *debate environment* — the parameters of a debate game that designers generally cannot change — then the *design elements* — those parameters that can be changed — and finally, the debate game. The motivation behind the *debate environment* and *design elements* will be clearer once the debate game is defined (Definition 2).

The debate environment

Definition 1 (Debate environment). *A debate environment is a tuple* $\mathbb{E} = \langle \mathcal{W}, \pi, \mathcal{Q}, \mathcal{A}, \tau, \mathcal{E} \rangle$ *which consists of:*
- *An arbitrary set $\mathcal{W}$ of **worlds** and a prior distribution $\pi \in \Delta(\mathcal{W})$ from which the current world is sampled.*
- *A set $\mathcal{Q}$ of **questions**, where each $q \in \mathcal{Q}$ is a text string.*
- *An arbitrary set $\mathcal{A}$ of **answers**.*
- *A mapping $\tau : \mathcal{Q} \times \mathcal{A} \times \mathcal{W} \to [0, \infty)$ which measures the **deviation** of an answer **from the truth** about the world.*
- *A set $\mathcal{E}$ of **experiments** $e : \mathcal{W} \to 2^{\mathcal{W}}$ the judge can perform to learn that the current world w belongs to $e(w)$.*

One example of a debate is a highly general case where $\mathcal{W}$ is the set of all the ways our environment might be, $q \in \mathcal{Q}$ is the set of questions we might ask, and $\mathcal{A}$ are the textual responses that AI debaters might produce. The mapping τ represents the deviation from the question's true answer, while experiments constitute a cheaper – and possibly less reliable – way of obtaining information. For the question *Which restaurant should I go to?*, $\tau(w, q, a)$ could indicate how dissatisfied one would be at each restaurant a, and an experiment could indicate my preference between two restaurants by comparing their menus.

We can also consider much more specific cases. For example, $\mathcal{W}$ may represent the set of all legal board positions in Go, and q asks "What is the optimal next move?"

For many kinds of questions, we may set $\tau(\cdot) = 0$ for all correct answers and $\tau(\cdot) = 1$ for incorrect or invalid ones.

The design elements of debate There are some rules for a debate that a designer can control, called the *design elements*:
- the choice of **question** q and **legal answers** $\mathcal{A}(q) \subset \mathcal{A}$,
- **communication language** $\mathcal{L}_c$ (an arbitrary set),
- **argumentation protocol** $P : \mathcal{Q} \times \mathcal{A}^2 \times \mathcal{L}_c^* \to 2^{\mathcal{L}_c}$,
- **termination** condition $\mathcal{T} \subset \mathcal{Q} \times \mathcal{A}^2 \times \mathcal{L}_c^*$,
- **experiment-selection** policy $E : \mathcal{T} \to \mathcal{E}$, and
- **debating incentives** $u_i : \mathcal{T} \times 2^{\mathcal{W}} \to [-1, 1]$.

In practice, some of these rules will be hard-coded, for example, a designer may restrict the answers to $\mathcal{A} = \{$"yes", "no"$\}$ or make it physically impossible for the debating agents to communicate in anything other than binary code ($\mathcal{L}_c = \{0, 1\}^*$). On the other hand, the designer may outsource the implementation of some rules to a human judge. For example, a designer can automatically prohibit repetition of a particular string, but the prohibition of rephrasing of previously made points needs to be delegated to the judge. Similarly, a debate can automatically terminate after N rounds, but a judge is needed to end after debaters no longer say anything relevant.[2]

Using the *debate environment* and *design elements*, a debate game can be formalized as follows:

Definition 2 (Debate and debate game). *A **debate** is a tuple $D = \langle \mathbb{E}, q, G \rangle$, where $\mathbb{E}$ is a debate environment, $q \in \mathcal{Q}$ is a question, and $G = (G_q)_{q \in \mathcal{Q}}$ is a **debate game**. Formally, each G_q is a two-player zero-sum extensive form[3] game that proceeds as follows:*

1. *The world w is sampled from π and shown to **debaters** 1 and 2 together with the question q.*
2. *The debaters simultaneously pick answers $a_1, a_2 \in \mathcal{A}(q)$.*
3. *The debaters alternate[4] making **arguments** $x_1, x_2, \dots \in \mathcal{L}_c$, where $x_j \in P(q, a_1, a_2, x_1, \dots, x_{j-1})$, stopping once $(q, a_1, a_2, x_1, x_2, \dots) = t \in \mathcal{T}$ is a **terminated dialogue**.*
4. *A single experiment $e = E(t)$ is selected and its result $e(w)$ is used as a context for the next step.*
5. *The debaters receive **utilities** $u_1(t, e(w)) \in [-1, 1]$ and $u_2(t, e(w)) = -u_1(t, e(w))$.*
6. *The answer of the debater with higher utility becomes the **outcome** $o(w, t)$ of D (with ties broken randomly).*

---

[1]A two-player zero-sum game is one where the utilities satisfy $u_2 = -u_1$. As a result, it suffices to consider the first player's utility u_1, and assume that player 2 is trying to minimize this number.

[2]We can distinguish between the designer of the debate, the person who selects the question q, and the judge who determines its winner, who may in-practice be different, or the same person. While the present text does not analyze the role of the human judge in detail, we believe that for such analysis, it is useful to view the judge not as a player in the debate game, but rather as some $J \in \mathcal{J}$ which parametrizes $\mathcal{A}(q) = \mathcal{A}^J(q)$, $P = P^J$, $\mathcal{T} = \mathcal{T}^J$, $E = E^J$, and $u_i = u_i^J$ (but *not* $\mathbb{E}$ and $\mathcal{L}_c$). We do, however, sometimes anthropomorphize parts of debate by speaking as if they were performed by a judge.

[3]EFGs are a standard model of sequential decision-making, described, for example, in (Osborne and Rubinstein, 1994). Partially-observable stochastic games (Hansen, Bernstein, and Zilberstein, 2004) constitute an equally valid (Kovařík et al., 2019) choice of model.

[4]That is, debater 1 makes odd arguments $x_1, x_3, \dots$ while 2 makes x_2, x_4, etc.

7. **Debate error** *is the resulting deviation* $\tau(q, w, o(w, t))$ *between the outcome of D and the true answer to q.*

The model makes several simplifications, but can be generalized in the following straightforward ways:

Remark 3 (Natural extensions of debate games)**.** *Debate games may be generalized with:*
- *Non-simultaneous answers. The roles in the answering phase might be asymmetric, in that one debater might see their opponent's answer before selecting their own.*[5]
- *Debaters with imperfect information. Instead of having perfect information about w, the debaters might only receive some imperfect observation $I_1(w)$, $I_2(w)$. This limitation is particularly relevant for scenarios involving human preferences, such as the restaurant example.*
- *Judge interaction. A judge that asks questions and makes other interjections may be added as a "chance" player J with fixed strategy.*
- *Dangerous experiments. In the real world, some experiments might have dangerous side-effects. This may be modelled by considering experiments $e : \mathcal{W} \to 2^{\mathcal{W}} \cup [-\infty, \infty]^3$ which sometimes just give information, but other times bypass the debate and assign utilities u_1, u_2 and the debate error directly.*
- *Generalized outcome-selection policies. Instead of always adopting the more-favoured, the judge may adopt an answer according to a mapping $o : \mathcal{T} \times 2^{\mathcal{W}} \to \mathcal{A}$, or may even be given an option of ignoring suspicious or uninformative debates.*

2.2 Properties of debate games

Debate phases and relation to game theory. For the purpose of modelling the debaters' actions, we distinguish between the **answering phase** (step 2 of Definition 2) and the **argumentation phase** (step 3). Once q, w, and (a_1, a_2) get fixed at the start of the argumentation phase, the utilities of the debaters only depend on the subsequent arguments x_j raised by the agents. Since the agents have full information about each argument, the *argumentation phase is a two-player zero-sum sequential game with perfect information*. We denote this game $G_{qwa_1a_2}$. To analyze the answering phase, we first recall an important property of two-player zero-sum games: In every such G', all Nash-equilibrium strategies σ^* result in the same expected utility $\mathbf{E}_{\sigma^*} u_1$ (Shoham and Leyton-Brown, 2008, Thm 3.4.4). This number is called the **value of the game** and denoted v^*. Assuming optimal argumentation strategies, each debater thus knows that playing the argumentation game $G_{qwa_1a_2}$ results in some utility $v^*_{qwa_1a_2}$. This allows them to abstract away the argumentation phase. By randomizing the order of argumentation and treating both debaters equally, we can further ensure that $v^*_{qwa_2a_1} = -v^*_{qwa_1a_2}$. As a result, each *answering phase is a symmetric two-player zero-sum matrix game* with actions $\mathcal{A}(q)$ and payoffs $v^*_{wqa_1a_2}$ (to player 1).

These observations have an important implication: Fully general EFGs might contain complications that make finding their solutions difficult. However, both the answering game

---

[5]As in, e.g., the Devil's advocate AI (Armstrong, 2017a).

and the argumentation game belong to highly-specific and well-understood subclasses of EFGs and are thus amenable to simpler solution techniques.

Measuring the usefulness of debate. We measure the suitability of a debate design by the degree to which optimal play by the debaters results in low debate-error. By default, we focus on the worst-case where both the world and the debate outcome are selected adversarially from the support of their respective distributions. We denote the *support* of a probability measure p by supp (p).

Definition 4 (Truth promotion)**.** *In the following, $D = \langle \mathbb{E}, q, G \rangle$ is a debate, $\epsilon \geq 0$, and w always denotes some element of $supp(\pi)$, σ a Nash-equilibrium strategy in G, and t a terminal dialogue compatible with σ in w. D is said to be:*
- *ϵ-truth promoting (in the worst-case) in w if for each σ, we have $\sup_t \tau(q, o(t, w), w) \leq \epsilon$,*
- *ϵ-truth promoting if it is ϵ-truth promoting in every w,*
- *and ϵ-truth promoting in expectation if for each σ, we have $\mathbf{E}_{w \sim \pi} \mathbf{E}_{t \sim \sigma} \tau(q, o(t, w), w) \leq \epsilon$.*

When a debate is 0-truth promoting, we refer to it simply as "truth-promoting". Finally, we formalize the *idealized* version of the design problem as follows:

Problem 5 (When is debate truth promoting?)**.** *For given $\langle \mathbb{E}, \cdot, G \rangle$, characterize those $q \in \mathcal{Q}$ for which any optimal strategy in $\langle \mathbb{E}, q, G \rangle$ only gives answers with $\tau(q, a, w) = 0$.*

3 Feature Debate

In order to explore the properties that debates can have, it is useful to have a toy version of the general framework. In this section, we discuss how questions can be represented as functions of "elementary features" of the world and describe a simple debate game in which the arguments are restricted to revealing these features. This is inspired by (Irving, Christiano, and Amodei, 2018), where each world is an image from the MNIST database, elementary features are pixels, and the question is "Which digit is this?". Rather than faithfully capturing all important aspects of debate, the purpose of the toy model is to provide a *simple* setting for investigating *some* aspects. The limitations of the model are further discussed in Section 6.

3.1 Defining Feature Debate

Questions about functions. Many questions are *naturally* expressed as enquiries q_f about some $f : \mathcal{W} \to \mathcal{X}$

$$q_f := \text{"What is output of } f\text{?"} \tag{1}$$

and come accompanied by the answer space $\mathcal{A} = \mathcal{X}$ (or $\mathcal{A} = \Delta(\mathcal{X})$). Examples include questions of measurement ("How far is the Moon?", "How much do I weigh?") and classification ("Which digit is on this picture?", "Will person A beat person B in a poker game?"). For simplicity, we focus on questions q_f about functions $f : \mathcal{W} \to [0, 1] = \mathcal{X}$ and the truth-mapping $\tau(q, w, a) := |f(w) - a|$.

These assumptions are not very restrictive — they include binary questions of the type "Is Y true?" ($\mathcal{X} = \{0, 1\}$), and their generalizations "How likely is Y to be true?" and "To what degree is Y true?". Any debate about

an "n-dimensional question" can be reduced into n "1-dimensional" debates, and any function $f : \mathcal{W} \to \mathbb{R}$ can be re-scaled to have range $[0, 1]$.

Feature Debate and Its Basic Properties. In feature debate, we assume that worlds are fully described by their **elementary features** — i.e. we suppose that $\mathcal{W} = \Pi_{i=1}^{\infty}[0, 1]$ and denote $W_i : w = (w_i)_{i=1}^{\infty} \in \mathcal{W} \mapsto w_i$. Moreover, we assume that each round consists of each debater making one argument of the form "the value of i-th feature W_i is x". We consider a judge who can experimentally verify any elementary feature (but no more than one per debate):

Definition 6 (Feature debate environment). *A **feature debate environment** $\mathbb{F}_\pi$ is a debate environment where:*
- *The prior distribution π is a (Borel) probability measure on $\mathcal{W} = [0, 1]^{\mathbb{N}}$.*
- *Each question pertains to a function f, i.e. $\mathcal{Q} = \{q_f \mid f : \mathcal{W} \to [0, 1]$ measurable$\}$.*
- *The answers are $\mathcal{A} = [0, 1]$.*
- *The deviation-from-truth is the distance $\tau(q_f, a, w) = |f(w) - a|$ between the answer a and the truth $f(w)$.*
- *Each experiment reveals one feature, i.e. $\mathcal{E} = \{e_i \mid i \in \mathbb{N}\}$, where $e_i(w) := \{\tilde{w} \mid \tilde{w}_i = w_i\}$.*

For full generality, we may want to assume the debaters can lie about the features, but for the analysis in this paper, we ignore this case. The reason is that if the opponent can point out a lie, and then the judge can test and penalize it, uttering this lie will be sub-optimal. We thus only consider truthful claims of the form "$W_i = w_i$". With a slight abuse of notation, this allows us to identify the communication language $\mathcal{L}_c$ with the feature-indexing set $\mathbb{N}$. Any argument sequence $\vec{i} := \vec{i}_k := (i_1, \ldots, i_k)$ thus effectively reveals the corresponding features. (We write "$W_{\vec{i}} = w_{\vec{i}}$".)

We suppose the judge has access to the world-distribution $\pi \in \mathcal{W}$ and can update it correctly on new information, but only has "patience" to process $2N$ pieces of evidence. Finally, the debaters are penalized for any deviation between their answer and the judge's final belief and — to make the debate zero-sum — are rewarded for any deviation of their opponent. Adopting the q_f shorthand introduced earlier, the formal definition is as follows:

Definition 7 (Feature debate). *A **feature debate** $F_\pi(f, N) = \langle \mathbb{F}_\pi, q_f, G \rangle$ is a debate with the following specific rules:*
- *A randomly selected player makes the first argument.*
- *$\mathcal{L}_c = \mathbb{N}$ and $P(f, a_1, a_2, i_1, \ldots, i_k) := \mathbb{N} \setminus \{i_1, \ldots, i_k\}$.*
- *After $2N$ arguments have been made, the judge sets $u_1(t, w_{\vec{i}}) := |\hat{f}(w_{\vec{i}}) - a_2| - |\hat{f}(w_{\vec{i}}) - a_1|$, where $\hat{f}(w_{\vec{i}})$ is the posterior mean*

$$\hat{f}(w_{\vec{i}}) := \mathbf{E}_\pi \left[f \mid W_{\vec{i}} = w_{\vec{i}} \right].^6$$

3.2 Optimal play in feature debate

The zero-sum assumption implies that any shift in $\hat{f}(w_{\vec{i}})$ will be endorsed by one debater and opposed by the other

---

(or both will be indifferent). The following symbols denote the two extreme values that the judge's final belief can take, depending on whether the debater who makes the first argument aims for high values of $\hat{f}(w_{\vec{i}})$ and the second debater aims for low ones ($\uparrow\downarrow$) or vice versa ($\downarrow\uparrow$):

$$\hat{f}_N^{\uparrow\downarrow}(w) := \max_{i_1} \min_{i_2} \ldots \max_{i_{2N-1}} \min_{i_{2N}} \hat{f}(w_{\vec{i}}),$$

$$\hat{f}_N^{\downarrow\uparrow}(w) := \min_{i_1} \max_{i_2} \ldots \min_{i_{2N-1}} \max_{i_{2N}} \hat{f}(w_{\vec{i}}).$$

Since $\max_x \min_y \varphi(x, y) \le \min_y \max_x \varphi(x, y)$ holds for any φ, the second debater always has an edge: $(\forall w \in \mathcal{W}) :$ $\hat{f}_N^{\uparrow\downarrow}(w) \le \hat{f}_N^{\downarrow\uparrow}(w)$. Lemma 8 (i) shows that if the order of argumentation is randomized as in Definition 7, the optimal answers lie precisely in these bounds. This result immediately yields a general error bound (ii) which will serve as an essential tool for further analysis of feature debate.

Lemma 8 (Optimal play in feature debate). *(i) The optimal answering strategies in $F_\pi(f, N)$ are precisely all those that select answers from the interval $[\hat{f}_N^{\uparrow\downarrow}(w), \hat{f}_N^{\downarrow\uparrow}(w)]$.*

(ii) In particular, $F_\pi(f, N)$ is precisely $\max\{|\hat{f}_N^{\uparrow\downarrow}(w) - f(w)|, |\hat{f}_N^{\downarrow\uparrow}(w) - f(w)|\}$-truth promoting in w.

4 When Do Feature Debates Track Truth?

In this section, we assess whether feature debates track truth under a range of assumptions.

4.1 Truth-Promotion and Critical Debate-Length

Some general debates might be so "biased" that no matter how many arguments an honest debater uses, they will not be able to convince the judge of their truth. Proposition 9 ensures that this is not the case in a typical feature debate:

Proposition 9 (Sufficient debate length). *$F_\pi(f, N)$ is truth-promoting for functions that depend on $\le N$ features.*

Proof. In an N-round debate about a question that depends on $\le N$ features, either of the players can unilaterally decide to reveal all relevant information, ensuring that $\hat{f}(w_{\vec{i}_{2N}}) = f(w)$. This implies that $\hat{f}_N^{\uparrow\downarrow}(w) = \hat{f}_N^{\downarrow\uparrow}(w) = f(w)$. The result then follows from Lemma 8. □

However, Proposition 9 is optimal in the sense that if the number of rounds is smaller than the number of critical features, the resulting debate error might be very high.

Proposition 10 (Necessary debate length). *When f depends on $N + 1$ features, the debate error in $F_\pi(f, N)$ can be 1 (i.e., maximally bad) in the worst-case world and equal to $\frac{1}{2}$ in expectation (even for continuous f).*

The "counterexample questions" which this result relies on are presented in the following Section 4.2. The formal proof and the continuous case are given in the appendix.

4.2 Three Kinds of Very Difficult Questions

We now construct three classes of questions which cause debate to perform especially poorly[7], in ways that are analogous to failures of realistic debates.

---

[6]That is, the judge considers the arguments to be generated independently, using a "naive-Bayes"-like assumption that ignores the fact that evidence may be selected strategically.

[7]While we focus on results in worst-case worlds, the illustrated behaviour might become the norm with a biased judge (Sec. 6.2).

Unfair questions. A question may be difficult to debate when **arguing for one side requires more complex arguments**. Indeed, consider a feature debate in a world w uniformly sampled from Boolean-featured worlds $\Pi_{i\in\mathbb{N}}W_i = \{0,1\}^{\mathbb{N}}$, and suppose the debate asks about the conjunctive function $\varphi := W_1 \wedge \ldots \wedge W_K$ for some $K \in \mathbb{N}$. In worlds with $w_1 = \cdots = w_K = 1$, an honest debater has to reveal K features to prove that $\varphi(w) = 1$. On the other hand, a debater arguing for the false answer $a = 0$ merely needs to avoid helping their opponent by revealing the features $W_1, \ldots, W_K$. In particular, this setup shows that a debate might indeed require as many rounds as there are relevant features, thus proving the worst-case part of Proposition 10).

Unstable debates. Even if a question does not bias the debate against the true answer as above, the **debate outcome** might still be **uncertain until the very end**. One way this could happen is if the judge always feels that more information is required to get the answer right. Alternatively, every new argument might come as a surprise to the judge, and be so persuasive that the judge ends up always taking the side of whichever debater spoke more recently.

To see how this behavior can arise in our model, consider the function $\psi := \text{xor}(W_1, \ldots, W_K)$ defined on worlds with Boolean features, and the world $w = (1, 1, \ldots)$.[8] If the world distribution π is uniform over $\{0,1\}^{\mathbb{N}}$, the judge will reason that no matter what the debaters say, the last unrevealed feature from the set $\{W_1, \ldots, W_K\}$ always has an equal chance of flipping the value of ψ and keeping it the same, resulting in $\hat{\psi}(w_{\vec{i}}) = \frac{1}{2}$. As a result, the only optimal way of playing $F_\pi(\psi, N)$ is to give the wrong answer $a = \frac{1}{2}$, unless a single debater can, by themselves, reveal all features $W_1, \ldots, W_K$. This happens precisely when $K \le N$. In particular, the case where $K = N + 1$ proves the "in expectation" part of Proposition 10.

To achieve the "always surprised and oscillating" pattern, we consider a prior π under which each each feature w_i is sampled independently from $\{0,1\}$, but in a way that is skewed towards $W_i = 0$ (e.g., $\Pr[W_i = 0] = 1 - \delta$ for some small $\delta > 0$). The result of this bias is that no matter which features get revealed, the judge will always be more likely to believe that "no more features with value $W_i = 1$ are coming" — in other words, the judge will be very confident in their belief while, at the same time, shifting this belief from 0 to 1 and back each round.

Distracting evidence. For some questions, there are misleading arguments that appear plausible and then require extensive counter-argumentation to be proven false. By making such arguments, a dishonest debater can stall the debate, until the judge "runs out of patience" and goes with their possibly-wrong surface impression of the topic. To illustrate the idea, consider the uniform distribution π over $\mathcal{W} = [0,1]^{\mathbb{N}}$ and a question q_f about some $f : \mathcal{W} \to [0,1]$ that only depends on the first K features. Suppose, for convenience of notation, that the debaters give answers $a_1 = 1$, $a_2 = 0$ and the sampled world is s.t. $f(w) = 1$ and

$w_{K+1} = w_{K+2} = \cdots = 1$. To adversarially modify f, we first define a function $S : [0,1]^2 \to [0,1]$ as $S(x,y) = 1$ if either $x \ne 1$ or $x = y = 1$ and as $S(x,y) = 0$ otherwise. By replacing f by $f'(w) := f(w)S(w_m, w_n)$, where $n > m > K$, S introduces an "unlikely problem" $x = 1$ and an "equally unlikely fix" $y = 1$, thus allowing the dishonest player 2 to "stall" for one round. Indeed, the presence of $S(\cdot, \cdot)$ doesn't initially affect the expected value of the function in any way. However, if player 2 reveals that $W_m = w_m = 1$, the expectation immediately drops to 0, forcing player 1 to "waste one round" by revealing that $W_n = w_n = 1$. To make matters worse yet, we could ask about $\hat{f}'(w) := f(w)\Pi_{i=1}^d S(w_{m_i}, w_{n_i})$, or use a more powerful stalling function $S(x, y_1 \wedge \cdots \wedge y_k)$ that forces the honest player to waste k rounds to "explain away" a single argument of the opponent.

4.3 Detecting Debate Failures

When a debater is certain that their opponent will not get a chance to react, they can get away with making much bolder claims.[9] The resulting "unfairness" is not a direct source for concern because the order of play can easily be randomized or made simultaneous. However, we may wish to measure the last-mover advantage in order to detect whether a debate is tracking truth as intended. The proof of Lemma 8 (in particular, equation (5)) yields the following result:

Corollary 11 (Last-mover advantage). *If optimal debaters in the feature debate* $F_\pi(f, N)$ *give answers* $a_1, a_2 \in [\hat{f}_N^{\uparrow\downarrow}(w), \hat{f}_N^{\downarrow\uparrow}(w)]$, *the debater who argues second will obtain* $|a_1 - a_2|$ *expected utility.*

Recall that, by Lemma 8, all answers from the interval $[\hat{f}_N^{\uparrow\downarrow}(w), \hat{f}_N^{\downarrow\uparrow}(w)]$ are optimal. Corollary 11 thus implies that even if the agents debate optimally, some portion of their utility – up to $\delta := \hat{f}_N^{\downarrow\uparrow}(w) - \hat{f}_N^{\uparrow\downarrow}(w)$ – depends on the randomized choice of argumentation order.[10] Incidentally, Lemma 8 implies that the smallest possible debate error is $\delta/2$ (which occurs when the true answer is $f(w) = (\hat{f}_N^{\uparrow\downarrow}(w) - \hat{f}_N^{\downarrow\uparrow}(w))/2$). This relationship justifies a simple, common-sense heuristic: If the utility difference caused by reversing the argumentation order is significant, our debate is probably far from truth-promoting.

5 Two Important Special Cases of Debate

As a general principle, a narrower class of debates might allow for more detailed (and possibly stronger) guarantees. We describe two such sub-classes of *general* debates and illustrate their properties on variants of feature debate.

5.1 Debate with Independent Evidence

When evaluating solution proposals in practice, we sometimes end up weighing its "pros" and "cons". In a way,

---

[9]Conversely, some realistic debates might provide first-mover advantage due to anchoring and framing effects.

[10]For illustration, a (literally) extreme case of last-mover advantage occurs in the "oscillatory" debate from Section 4.2, where the interval $[\hat{f}_N^{\uparrow\downarrow}(w), \hat{f}_N^{\downarrow\uparrow}(w)]$ spans the whole answer space $[0,1]$.

we are viewing these arguments as (statistically) independent evidence related to the problem at hand. This is often a reasonable approximation, e.g., when deciding which car to buy, and sometimes an especially good one, e.g., when interpreting questionnaire results from different but independent respondents. We now show how to model these scenarios as feature debates with statistically independent features, and demonstrate the particularly favourable properties.

Feature debate with independent evidence. As a mathematical model of such setting, we consider $\mathcal{X} = \{0, 1\}$, $\mathcal{W} := \Pi_{i=1}^{\infty}[0, 1]$, and denote by $W_i : (w, x) \mapsto w_i \in [0, 1]$ and $X : (w, x) \mapsto x \in \{0, 1\}$ the coordinate projections in $\mathcal{W} \times \mathcal{X}$. Informally, we view the last coordinate as an unknown feature of the world and the debate we construct will be asking "What is the value of this unknown feature?". To enable inference about X, we consider some probability distribution $\mathbb{P}$ on $\mathcal{W} \times \mathcal{X}$. (For convenience, we assume $\mathbb{P}$ is discrete.) Finally, to be able to treat arguments of the form "$W_i = w_i$" as independent evidence related to X, we assume that the features W_i, $i \in \mathbb{N}$, are mutually independent conditional on X.[11] To describe this setting as a feature-debate, we define π as the marginalization of $\mathbb{P}$ onto $\mathcal{W}$ and consider the question $q =$ "How likely is $x = 1$ in our world?", i.e. "what is the value of f, where $f(w) := \mathbf{E}_{\mathbb{P}}[X \mid \mathcal{W} = w]$". We denote the resulting "independent evidence" feature debate $F_\pi(q, N)$ as $F^{\text{ie}}(\mathbb{P}, N)$.

Judge's belief and its updates. Firstly, recall that any probability can be represented using its odds form, which is equivalent to the corresponding log-odds form:

$$\mathbb{P}(A) \in [0, 1] \longleftrightarrow \mathbb{P}(A)/\mathbb{P}(\neg A) \in [0, \infty]$$
$$\longleftrightarrow \log(\mathbb{P}(A)/\mathbb{P}(\neg A)) \in [-\infty, \infty].$$

Moreover, when expressed in the log-odds form, Bayes' rule states that updating one's belief in hypothesis H in light of evidence A is equivalent to shifting the log-odds of the prior belief by $\log(\mathbb{P}(A \mid H) / \mathbb{P}(A \mid \neg H))$.

At any point in the debate $F^{\text{ie}}(\mathbb{P}, N)$, the judge's belief $\hat{f}(w_{\vec{i}}) = \mathbf{E}[f \mid W_{\vec{i}} = w]$ is, by the definition of f, equal to the conditional probability $P(X = 1 \mid W_{\vec{i}} = w_{\vec{i}})$. To see how the belief develops over the course of the debate, denote by $\beta_{\vec{i}}(w)$ the corresponding log-odds form. Initially, $\hat{f}_\emptyset(w)$ is equal to the prior $\mathbb{P}(X = 1)$, which corresponds to $\beta_\emptyset(w) = \log(\mathbb{P}(X = 1)/\mathbb{P}(X = 0)) =: p_0$. Denoting

$$\text{ev}_j(w) := \log\left(\frac{\mathbb{P}(W_j = w_j \mid X = 1)}{\mathbb{P}(W_j = w_j \mid X = 0)}\right),$$

the above form of the Bayes' rule implies that upon hearing an argument "$W_j = w_j$", the judge will update their belief according to the formula $\beta_{\vec{i},j}(w) = \beta_{\vec{i}}(w) + \text{ev}_j(w)$. In other words, the arguments in $F^{\text{ie}}(\mathbb{P}, N)$ combine additively:

$$\beta_{\vec{i}_n}(w) = p_0 + \text{ev}_{i_1}(w) + \cdots + \text{ev}_{i_n}(w). \quad (2)$$

---

[11]In other words, $\mathbb{P}(W_j = w_j \mid X = x)$ is equal to $\mathbb{P}(W_j = w_j \mid X = x, W_{\vec{i}_k} = w_{\vec{i}_k})$ for every x, w_j, and $w_{\vec{i}_k}$.

Optimal strategies and evidence strength. Recall that positive (negative) log-odds correspond to probabilities closer to 1 (resp. 0). Equation (2) thus suggests that for any w, the arguments $\mathbb{N}$ can be split into three "piles" from which the debaters select arguments: $\mathcal{N}_\uparrow := \{j \in \mathbb{N} \mid \text{ev}_j(w) > 0\}$ containing arguments supporting the answer "$X = 1$ with probability 100%", the pile $\mathcal{N}_\downarrow = \{j \in \mathbb{N} \mid \text{ev}_j(w) < 0\}$ of arguments in favor of the opposite, and the irrelevant arguments $\mathcal{N}_{\text{ir}} = \{j \in \mathbb{N} \mid \text{ev}_j(w) = 0\}$. As long as the debaters give different answers, one of them will use arguments from $\mathcal{N}_\downarrow$, while the other will only use $\mathcal{N}_\uparrow(w)$ (both potentially falling back to $\mathcal{N}_{\text{ir}}$ if their pile runs out).[12] Moreover, a rational debater will always use the strongest arguments from their pile, i.e. those with the highest **evidence strength** $|\text{ev}_j(w)|$. Correspondingly, we denote the "total evidence strength" that a players can muster in n rounds as

$$\text{Ev}_n^\uparrow(w) := \max\left\{\sum_{j \in J} \text{ev}_{i_k}(w) \mid J \subset \mathbb{N}, |J| = n\right\} \text{ and}$$

$$\text{Ev}_n^\downarrow(w) := \max\left\{\sum_{j \in J} (-\text{ev}_{i_k}(w)) \mid J \subset \mathbb{N}, |J| = n\right\}.$$

(To make the discussion meaningful, we assume the evidence sequence $(\text{ev}_j(w))_j$ is bounded and the maxima above are well-defined.) The equation (2) implies that – among optimal debaters – one always selects arguments corresponding to $\text{Ev}_N^\uparrow(w)$ while the other aims for $\text{Ev}_N^\downarrow(w)$. Since this holds independently of the argumentation order, we get $\hat{f}_N^{\uparrow\downarrow}(w) = \hat{f}_N^{\downarrow\uparrow}(w)$. Together with Lemma 8, this observation yields the following result:

Corollary 12 (Unique optimal answer)**.** *In the answering phase of any $F^{\text{ie}}(\mathbb{P}, N)$ with bounded evidence, the only Nash equilibrium is to select the $\hat{f}_N^*(w)$ which satisfies*

$$\hat{f}_N^{\uparrow\downarrow}(w) = \hat{f}_N^{\downarrow\uparrow}(w) = \hat{f}_N^*(w) := \text{the probability}$$

corresponding to the log-odds $p_0 + Ev_N^\uparrow(w) - Ev_N^\downarrow(w)$.

Debate error. To compute the debate error in $F^{\text{ie}}(\mathbb{P}, \cdot)$, denote the strength of the evidence that remains in each debater's "evidence pile" after n rounds as

$$R_n^\uparrow(w) := \sum_{j \in \mathcal{N}_\uparrow} \text{ev}_j(w) - \text{Ev}_n^\uparrow(w), \quad (3)$$

$$R_n^\downarrow(w) := \sum_{j \in \mathcal{N}_\downarrow} (-\text{ev}_j(w)) - \text{Ev}_n^\downarrow(w). \quad (4)$$

Furthermore, assume that additionally to $(\text{ev}_j(w))_j$ being bounded, the total evidence $\lim_{n \to \infty} \text{Ev}_n^a(w)$ in favor of a is infinite for at most one $a \in \{\uparrow, \downarrow\}$.[13] Since the true answer $f(w) = \mathbb{P}(X = 1 \mid \mathcal{W} = w)$ corresponds to $p_0 + \sum_{i \in \mathbb{N}} \text{ev}_j(w) = p_0 + \sum_{j \in \mathcal{N}_\uparrow} \text{ev}_j(w) - \sum_{j \in \mathcal{N}_\downarrow} (-\text{ev}_j(w))$, *the difference between the (log-odds forms of) the judge's final belief and the optimal answer is $R_N^\uparrow(w) - R_N^\downarrow(w)$.*

---

[12]Formally, these argumentation incentives follow from the first paragraph in the proof of Lemma 8.

[13]Note that the intuition "$\lim_n R_n^{(\cdot)}(w) = 0$" only fits if either $\mathcal{N}_a$ or $\mathcal{N}_{\text{ie}}$ is infinite. If a debater eventually has to reveal evidence against their own case, the numbers $R_n^{(\cdot)}(w)$ will get negative.

Early stopping and online estimation of the debate error. If we further assume that the debaters reveal the strongest pieces of evidence first, we can predict a debate's outcome before all N rounds have passed. If the n-th argument of player p has strength $|\text{ev}_i(w)| =: s_{p,n}(w)$, we know that further $N - n$ rounds of debate cannot reveal more than $(N - n)s_{p,n}(w)$ evidence in favor of p. This implies that as soon as the currently-losing player is no longer able to shift the judge's belief beyond the midpoint between the initial answers, we can stop the debate without affecting its outcome. If we further know that the question at hand depends on K features or less, we can also bound the difference between $f(w)$ and $\hat{f}(w_{\vec{i}})$. Indeed, in the worst-case scenario, all remaining arguments were all in favor of the same player p — even in this case, the (log-odds form of) $f(w)$ can be no further than $\max_{p=1,2} s_{p,N}(w)(K - 2N)$ away from the log-odds form of the final belief $\hat{f}(w_{\vec{i}_{2N}})$.

5.2 Debate with Information-Limited Arguments

Sometimes, a single argument cannot convey all relevant information about a given feature of the world. For example, we might learn that a person A lives in a city B, but not their full address, or – in the language of feature debate – learn that w_i lies in the interval $[0.5, 1]$, rather than understanding right away that $w_i = 0.75$. In such cases, it becomes crucial to model the judge's information bandwidth as limited.

Feature Debate Representation. In feature debate, we can represent each elementary feature $w_i \in [0, 1]$ in its binary form (e.g., $(0.75)_2 = 0.11000\ldots$), and correspondingly assume that each argument reveals one bit of some w_i. More specifically, we assume that (a) the debaters make arguments of the form "the n-th bit of i-th feature has value b", (b) they have to reveal the n-th bit of w_i before its $(n+1)$-th bit, and – using the same argument as in feature debate – (c) their claims are always truthful. Informally, each argument in this "information-limited" feature[14] debate $F_\pi^l(f, N)$ thus corresponds to selecting a dimension $i \in \mathbb{N}$ and doing a "50% zoom" on w along this dimension (Figure 1).

Performance bounds. By offering intermediate steps between features being completely unknown and fully revealed, the debate $F^{\text{ie}}(\cdot)$ allows for more nuanced guaran-

---

[14]The name is justified since $\mathbb{N}^2$ is isomorphic to $\mathbb{N}$ and thus $F_\pi^l(f, N)$ is formally equivalent to some feature debate $F_{\bar{\pi}}(\tilde{f}, N)$.
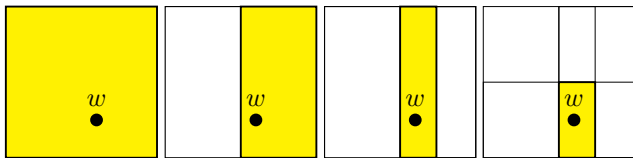

Figure 1: Each argument in an information-limited debate reduces the set of feasible worlds by "zooming-in" on the sampled world $w = (0.6, 0.25)$ along one dimension of $\mathcal{W}$. Here, the first two arguments provide information about w_1 (the x-axis) and the third one about w_2.

tees than those from Section 4.1. (Informally stated, the assumptions of the Proposition 13 can be read as "the values of f differ by at most L across different worlds, with feature w_i being responsible for up to a $\frac{1}{2^i}$-fraction[15] of the variance".)

Proposition 13. *Suppose that $f : \mathcal{W} \to \mathbb{R}$ is L-Lipschitz continuous[16] w.r.t. the metric $\rho(w, w') = \sum_{i \in \mathbb{N}} 2^{-i}|w_i - w_i'|$ on $\mathcal{W}$. Then $F_\pi^l(f, N)$ is $L/2^{\lfloor \sqrt{N} \rfloor}$-truth promoting.*

In contrast to Proposition 9, the L-Lipschitz assumption thus allows us to get approximately correct debate outcomes long before having full knowledge of all features. Note that the importance of Proposition 13 is not in the particular choice of weights, but rather in showing that argument weights can be translated into debate error bounds.

6 Limitations and Future Work

The language introduced so far does not fully capture all aspects of realistic AI debates — due to space limitations, it is simply not possible to cover all design variants and emergent phenomena in this initial work. In this section, we outline some notable avenues for making the debate model more accurate and useful, either by using an alternative instantiation of the general framework from Section 2 or by extending the toy model from Section 3. We start by discussing the modifications which are likely to improve the debate's performance or applicability, and follow-up by those which could introduce new challenges. For suggested future work on AI debate that is not specific to *modelling*, we refer the reader to Irving, Christiano, and Amodei (2018).

6.1 Plausible Improvements to Debate

Commitments and high-level claims. As Irving, Christiano, and Amodei (2018) suggest, an important reason why debate might work is that the debaters can make abstract or high-level claims that can potentially be falsified in the course of the debate. For example, debaters might start out disagreeing whether a given image — of which the judge can only inspect a single pixel — depicts a dog or a cat. Debater 1 might then claim that "here in the middle, there is a brown patch that is the dog's ear", to which their opponent counters "the brown patch is a collar on an otherwise white cat". Such exchanges would continue until one debater makes a claim that is (i) inconsistent with their answer, (ii) inconsistent with "commitments" created by previous arguments, or (iii) specific enough to be verified by the judge. In this example, (iii) might arise with an exchange "this pixel is white, which could not happen if it belonged to a brown dog's ear", "actually, the pixel is brown", which allows the judge to determine the winner by inspecting the pixel.

This ability to make high-level claims and create new commitments will often make the debate more time-efficient

---

[15]While similar results hold for general "weight ratios" between features, we chose weights $\frac{1}{2^i}$ for their notational convenience.

[16]A function is Lipschitz continuous with constant $L \geq 0$ (w.r.t. a metric ϱ) if it satisfies $|f(x) - f(y)| \leq L\varrho(x, y)$.

and incentivize consistency. Since consistency should typically advantage debaters that describe the true state of the world, commitments and high-level claims seem critical for the success of debate. We thus need a communication language $\mathcal{L}_c$ that is rich enough to enable more abstract arguments and a set of **effect rules** (Prakken, 2006) which specify how new arguments affect the debaters' commitments. To reason about such debates, we further need a model which relates the different commitments, to arguments, initial answers, and each other.

One way to get such a model is to view $\mathcal{W}$ as the set of assignments for a Bayesian network. In such setting, each question $q \in \mathcal{Q}$ would ask about the value of some node in $\mathcal{W}$, arguments would correspond to claims about node values, and their connections would be represented through the structure of the network. Such a model seems highly structured, amenable to theoretical analysis, and, in the authors' opinion, intuitive. It is, however, not necessarily useful for practical implementations of debate, since Bayes networks are computationally expensive and difficult to obtain.

Detecting misbehaviour. One possible failure of debate is the occurrence of stalling, manipulation, collusion, or selective provision of evidence. To remedy these issues, we can introduce specific countermeasures for these strategies. One option for quantifying the contribution of discourse to a human's understanding is to measure the changes in their ability to pass "exams" (Armstrong, 2017b). Another countermeasure would be to instantiate a meta-debate on the question of whether a debater is arguing fairly. However, such a meta-debate may, in some cases be even more challenging to judge correctly than the original question.

Alternative utility functions. We have considered utility functions that are linear in each debater's deviation $\Delta_p := |\hat{f}(w_{\vec{i}}) - a_p|$ from the judge's belief. However, other approaches such as "divide the total reward in proportion to $\frac{1}{\Delta_p}$" might give different but interesting results.

Real-world correspondence. To learn which real-world debates are useful on the one hand, and which theoretical issues to address on the other, a better understanding of the correspondence between abstract debate models and real-world debates is needed. For example, which real-world debates can be modelled as having independent evidence, being Lipschitz, having distracting arguments, and so on?

6.2 Obstacles in Realistic Debates

Sub-optimal judges. Some debates might have a canonical idealized way of being judged, which the actual judge deviates from at some steps. A fruitful avenue for future research is to investigate the extent to which debate fails gracefully as the judge deviates from this ideal. For example, games are canonically judged as giving a score of 1 to the winner and 0 to the loser. We could thus measure how much (and in what ways) the utility function can be modified before the game's winner is changed. Another approach would be to consider a judge that was biased. An unbiased Bayesian judge would set the prior to the true world-distribution, update the prior on each revealed piece of ev-

idence, and, at the end of the debate, calculate the corresponding expectation over answers. To model a judge who performs the first step imperfectly, we could consider a biased prior $\tilde{\pi} \in \Delta \mathcal{W}$ (distinct from the true world distribution π) and calculate the utilities using the corresponding biased belief $\tilde{f}(w_{\vec{i}}) := \mathbf{E}_{\tilde{\pi}} \left[f \mid W_{\vec{i}} = w_{\vec{i}} \right]$.

Manipulation. So far, we have described failures that come from "the judge being imperfect in predictable ways". However, real-world debates might also give rise to undesirable argumentation strategies inconceivable in the corresponding simplified model. For example, a debater might learn to exploit a bug in the implementation of the debate or, analogously, find a "bug" in the human judge. Worse yet, debaters might attempt to manipulate the judge using bribery or coercion. Note that for such tactics to work, the debaters need not be able to carry out their promises and threats — it is merely required that the judge believes they can.

Collusion. Without the assumption of zero-sum rewards, the debaters gain incentives to collaborate, possibly at the expense of the accuracy of the debate. Such "I won't tell on you if you don't tell on me" incentives might arise, for example, if both agents are given a positive reward if both answers seem good (or negative reward when the debate becomes inconclusive).

Sub-optimal debaters. If debaters argue sub-optimally, we might see new types of fallacious arguments. We should also expect to see stronger debaters win even in situations that advantage their weaker opponent. There could also be cases where the losing player complicates the debate game on purpose to increase variance in the outcome. Both of these phenomena can be observed between humans in games like Go, and we should expect analogous phenomena in general AI debate. One way of modelling asymmetric capabilities is to let two debaters run the same debating algorithm with a different computational budget (e.g., Monte Carlo tree search with a different number of rollouts).

7 Related Work

AI safety via debate. The kind of debate we sought to model was introduced in (Irving, Christiano, and Amodei, 2018), wherein it was proposed as a scheme for safely receiving advice from highly capable AI systems. In the same work, Irving et al. carried out debate experiments on the MNIST dataset and proposed a promising analogy between AI debates and the complexity class PSPACE. We believe this analogy can be made compatible with the framework introduced in our Section 2, and deserves further attention. Kovařík (2019) then demonstrated how to use debate to train an image classifier and described the design elements of debate in more detail. AI debate is closely related to two other proposals: (1) "factored cognition" (Ought, 2018), in which a human decomposes a difficult task into subtasks, each of which they can solve in a manageable time (similarly to how debate eventually zooms in on an easily-verifiable claim), and (2) "iterated distillation and amplification" (Christiano, Shlegeris, and Amodei, 2018), in which

a "baseline human judgement" is automated and amplified, similarly to how AI debates might be automated.

Previous works on argumentation. Persuasion and argumentation have been extensively studied in areas such as logic, computer science, and law. The introduction by Prakken (2006) describes a language particularly suitable for our purposes. Conversely, the extensive literature on argumentation frameworks (Dung, 1995) seems less relevant. The main reasons are (i) its focus on non-monotonic reasoning (where it is possible to retract claims) and (ii) that it assumes the debate language and argument structure as given, whereas we wish to study the connection between arguments and an underlying world model. AI systems are also being trained to identify convincing *natural-language* arguments — for a recent example, see, e.g., Perez et al. (2019).

Zero-sum games. As noted in the introduction, we can view two-player zero-sum games as a debate that aims to identify the game's winner (or an optimal strategy). Such games thus serve as a prime example of a problem for which the state of the art approach is (interpretable as) debate (Silver et al., 2017). Admittedly, only a small number of problems are formulated as two-player zero-sum games *by default*. However, some problems can be reformulated as such games: While it is currently unclear how widely applicable such "problem gamification" is, it has been used for combinatorial problems (Xu and Lieberherr, 2019) and theorem proving (Japaridze, 2009). Together with Silver et al. (2017), these examples give some evidence that the AI debate might be competitive (with other problem-solving approaches) for a wider range of tasks.

8 Conclusion

We have introduced a general framework for modelling AI debates that aim to amplify the capabilities of their judge and formalized the problem of designing debates that promote accurate answers. We described and investigated "feature debate", an instance of the general framework where the debaters can only make statements about "elementary features" of the world. In particular, we showed that if the debaters have enough time to make all relevant arguments, feature debates promote truth. We gave examples of two subclasses of debate: those where the arguments provide statistically independent evidence about the answers and those where the importance of different arguments is bounded in a known manner. We have shown that feature debates belonging to these sub-classes are approximately truth-promoting long before having had time to converge fully. However, we also identified some feature-debate questions that incentivize undesirable behaviour such as stalling, confusing the judge, or exploiting the judge's biases, resulting in debates that are unfair, unstable, and generally insufficiently truth-promoting. Despite its simplicity, feature debate thus allows for modelling phenomena that are highly relevant to issues we expect to encounter in realistic debates. Moreover, its simplicity makes feature debate well-suited for the initial exploration of problems with debate and testing of the corresponding solution proposals. Finally, we outlined multiple ways in which our model could be made more realistic —

among these, allowing debaters to make high-level claims seems like an especially promising avenue.

9 Acknowledgements

References

Armstrong, S. 2017a. Like this world, but... AI Alignment Forum.

Armstrong, S. 2017b. True understanding comes from passing exams. AI Alignment Forum.

Christiano, P.; Shlegeris, B.; and Amodei, D. 2018. Supervising strong learners by amplifying weak experts. *arXiv preprint arXiv:1810.08575*.

Dung, P. M. 1995. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial intelligence* 77(2):321–357.

Hansen, E. A.; Bernstein, D. S.; and Zilberstein, S. 2004. Dynamic programming for partially observable stochastic games. In *AAAI*, volume 4, 709–715.

Irving, G.; Christiano, P.; and Amodei, D. 2018. AI safety via debate. *arXiv preprint arXiv:1805.00899*.

Japaridze, G. 2009. In the beginning was game semantics? In *Games: Unifying logic, language, and philosophy*. Springer. 249–350.

Kovařík, V.; Schmid, M.; Burch, N.; Bowling, M.; and Lisỳ, V. 2019. Rethinking formal models of partially observable multiagent decision making. *arXiv preprint arXiv:1906.11110*.

Kovařík, V. 2019. AI safety debate and its applications. AI Alignment Forum.

Osborne, M. J., and Rubinstein, A. 1994. *A course in game theory*. MIT press.

Ought. 2018. Factored cognition. https://ought.org/research/factored-cognition. Accessed: 2019-09-25.

Perez, E.; Karamcheti, S.; Fergus, R.; Weston, J.; Kiela, D.; and Cho, K. 2019. Finding generalizable evidence by learning to convince Q&A models. In *Proceedings of EMNLP-IJCNLP 2019*, 2402–2411.

Prakken, H. 2006. Formal systems for persuasion dialogue. *The knowledge engineering review* 21(2):163–188.

Shoham, Y., and Leyton-Brown, K. 2008. *Multiagent systems: Algorithmic, game-theoretic, and logical foundations*. Cambridge University Press.

Silver, D.; Schrittwieser, J.; Simonyan, K.; Antonoglou, I.; Huang, A.; Guez, A.; Hubert, T.; Baker, L.; Lai, M.; Bolton, A.; et al. 2017. Mastering the game of go without human knowledge. *Nature* 550(7676):354.

Xu, R., and Lieberherr, K. 2019. Learning self-game-play agents for combinatorial optimization problems. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, 2276–2278.

A Proofs

We now give the full proofs of the results from the main text.

Proof of Lemma 8. Fix π, f, w, and N, and denote $\Lambda := [\hat{f}^{\uparrow\downarrow}, \hat{f}^{\downarrow\uparrow}] := [\hat{f}_N^{\uparrow\downarrow}(w), \hat{f}_N^{\downarrow\uparrow}(w)]$. By definition of $u_1(t) := |\hat{f}(w_{\vec{i}}) - a_2| - |\hat{f}(w_{\vec{i}}) - a_1|$, the utility u_1 is non-decreasing in $\hat{f}(w_{\vec{i}})$ when $a_1 \geq a_2$ (and in turn, u_2 is non-increasing in $\hat{f}(w_{\vec{i}})$). It follows that *an* optimal argumentation strategy is for player 1 to maximize $\hat{f}(w_{\vec{i}})$ and for player 2 to minimize it (and vice versa when $a_1 \leq a_2$). When calculating utilities, we can, therefore, assume that the final belief is either $\hat{f}(w_{\vec{i}}) = \hat{f}^{\uparrow\downarrow}$ or $\hat{f}(w_{\vec{i}}) = \hat{f}^{\downarrow\uparrow}$, depending on whether the player who argues second gave the higher or lower answer.

Suppose the players gave answers $\{a, b\}$, and the one arguing for b goes second. In this scenario, denote by $v^*(a, b)$ the utility this player receives if both argument optimally. By applying the above observation (about optimal argumentation strategies) in all possible relative positions of a, b, and $\hat{f}^{\uparrow\downarrow} \leq \hat{f}^{\downarrow\uparrow}$, we deduce that

$$v^*(a, b) = |a - b| - \text{dist}(b, \Lambda). \tag{5}$$

Denote by $v_1^*(a_1, a_2) := \frac{1}{2} v^*(a_2, a_1) - \frac{1}{2} v^*(a_1, a_2)$ the expected utility of player 1 if answers a_1 and a_2 are given (by players 1 and 2) and players argument optimally (the expectation is w.r.t. the randomized argumentation order). Using the formula for $v^*(a, b)$, we get

$$v_1^*(a_1, a_2) = \frac{1}{2} v^*(a_2, a_1) - \frac{1}{2} v^*(a_1, a_2)$$
$$= \frac{1}{2}(|a - b| - \text{dist}(a_1, \Lambda)) - \frac{1}{2}(|a - b| - \text{dist}(a_2, \Lambda))$$
$$= \frac{1}{2} \text{dist}(a_2, \Lambda) - \frac{1}{2} \text{dist}(a_1, \Lambda).$$

It follows that, independently of what strategy the opponent uses, it is always beneficial to give answers from within Λ (and that within this interval, all answers are equally good in expectation).[17] □

Proof of Proposition 10. For the worst-case part, suppose that π is a uniform distribution over $[0, 1]^{\mathbb{N}}$, $w = 1$, and let $\varphi := W_1 \wedge \cdots \wedge W_{N+1}$ be as in Sec. 4.2. In $F_\pi(\varphi, N)$, we have $\hat{\varphi}(w_{\vec{i}}) = 1$ if $\vec{i} \supset \{1, \ldots, N+1\}$ and $\hat{\varphi}(w_{\vec{i}}) = 0$ otherwise. In particular, we have $\hat{\varphi}_N^{\uparrow\downarrow}(w) = \hat{\varphi}_N^{\downarrow\uparrow}(w) = 0$, since the minimizing player can always select i_k from $\mathbb{N} \setminus \{1, \ldots, N+1\}$. Since $\varphi(w) = 1$, the result follows from Lemma 8.

---

[17]Incidentally, a more interesting behavior arises if the order of argumentation is fixed; in such a case, player 2 "flips a coin" between $\hat{f}^{\uparrow\downarrow}$ and $\hat{f}^{\downarrow\uparrow}$ and player 1 picks any of the strategies σ_1 with $\text{supp}(\sigma_1) \subset \Lambda$ and $\mathbf{E}[a_1 \mid a_1 \sim \sigma_1] = \frac{1}{2}(\hat{f}^{\uparrow\downarrow} + \hat{f}^{\downarrow\uparrow})$.

For the "in expectation" part, suppose that π is a uniform distribution over $\{0, 1\}^{N+1}$ (where each $w \in \{0, 1\}^{N+1}$ is extended by an infinite sequence of zeros) and $\psi(w) := \text{xor}(w_1, \ldots, w_{N+1})$. Recall that the value of $\psi(w)$ will be 1 if the total number of w_i-s with $w_i = 1$ is odd and 0 when the number is even. Until all of the $N + 1$ features of any w are revealed, there will be a 50% prior probability of an odd number of them having value 1 and 50% prior probability of an even number of them having value 1. As a result, we have $\hat{\psi}(w_{\vec{i}}) = \frac{1}{2}$ unless $\vec{i} \supset \{1, \ldots, N+1\}$. By not revealing any of these features in $F_\pi(\psi, N)$, either of the players can thus achieve $\hat{\psi}(w_{\vec{i}}) = \frac{1}{2}$. It follows that $\hat{\psi}_N^{\uparrow\downarrow}(w) = \hat{\psi}_N^{\downarrow\uparrow}(w) = \frac{1}{2}$, the only optimal strategy is to give the answer $a = \frac{1}{2}$, and — since f takes on only values 0 and 1 – the expected debate error is $\frac{1}{2}$.

Our counterexample function φ was discontinuous. However, the same result could be achieved by using $f(w) := \Pi_{i=1}^{N+1} w_i$ together with the uniform prior *over* $\{0, 1\}^{\mathbb{N}}$ — this yields a function that is continuous over its (discrete) domain. Alternatively, we could use the uniform prior over $[0, 1]^{\mathbb{N}}$ together with the continuous approximation $\Pi_{i=1}^{N+1} w_i^K$ (for some large $K \in \mathbb{N}$) of φ. Similarly, we could find a continuous approximation of ψ. This observation concludes the proof of the last part of the proposition. □

Proof of Proposition 13. Let $w \in \mathcal{W}$. For $\vec{i}$, denote by $\mathcal{W}(\vec{i})$ the set of worlds w' would reveal the same bits as w under the argument sequence $\vec{i}$. To show that the debate is ϵ-truth promoting in w, it suffices, by Lemma 8, to show that

$$f(w) - \epsilon \leq \hat{f}_N^{\uparrow\downarrow}(w) \text{ and } \hat{f}_N^{\downarrow\uparrow}(w) \leq f(w) + \epsilon.$$

To get the second inequality, recall that $\hat{f}_N^{\downarrow\uparrow}(w)$ is defined as the expected value of f under the judge's belief after N rounds of debate, under the assumption that player 1 attempts to drive the expectation as low as possible and player 2 as high as possible. Suppose that the player who is the first to argue chooses his arguments $i_1, i_3, i_5, \ldots$ in a way that minimizes the diameter $\max\{\rho(w, w') \mid w' \in \mathcal{W}(\vec{i})\}$ of $\mathcal{W}(\vec{i})$, i.e. by following the sequence $(1, 1, 2, 1, 2, 3, 1, \ldots)$. A simple calculation shows that after $N = 1 + \cdots + n$ rounds, this necessarily gets the diameter of $\mathcal{W}(\vec{i})$ to $\frac{n+2}{2^{n+1}}$ or less (depending on the opponent's actions $i_2, i_4, \ldots$). Since f is L-Lipschitz, we get that $f(w') \leq f(w) + L \cdot \frac{n+2}{2^{n+1}}$ on $\mathcal{W}(\vec{i})$. Since the optimal strategy of minimizing $\hat{f}_N^{\downarrow\uparrow}(w)$ can perform no worse than this, we get $\hat{f}_N^{\downarrow\uparrow}(w) \leq f(w) + L \cdot \frac{n+2}{2^{n+1}}$.

We derive the first inequality analogously. To finish the proof, we need to rewrite the inequality in terms of N. We have $\frac{n+2}{2^{n+1}} \leq \frac{2n}{2^{n+1}} \leq \frac{1}{2^n}$. Since we have $N \geq 1 + \cdots + \lfloor \sqrt{N} \rfloor$ for each $N \in \mathbb{N}$, we get $\hat{f}_N^{\downarrow\uparrow}(w) \leq f(w) + L \cdot \frac{1}{2^{\lfloor \sqrt{N} \rfloor}}$, which concludes the proof. □