Deceit and Credibility in Autonomous Agents: *Press Diplomacy* as a Laboratory

Benjamin J. Radford
Department of Political Science and Public Administration
Public Policy Ph.D. Program
School of Data Science
University of North Carolina at Charlotte
Charlotte, NC, USA
benjamin.radford@uncc.edu

I. Introduction

Advances in natural language processing portend a proliferation of chat bots indistinguishable from humans. Transformer-based language models, like GPT-2 and GPT-3, produce long-form human-like text and can respond, with superficial intelligence, to specific prompts. As modestly-funded adversaries seek to leverage these new technologies to control, understand, and steer the information environment, we should expect to see autonomous agents capable of not only engaging in conversation, but autonomous agents capable of changing minds. This requires agents to reason about others' perceptions of their own interests and to craft compelling narratives that are consistent with those perceptions regardless of their veracity. In other words, agents will need to lie strategically in order to mislead convincingly. Similarly, autonomous (or semi-autonomous) agents operating in an adversarial information environment will need the capability to reason about their counterparts' interests in order to discriminate between information and disinformation. In other words, knowledge-based fact checking is not sufficient to detect all disinformation; disinformation related to one's intent is not something that can be checked against a database of facts but must instead be identified via reasoning about the disinformation in the context of the speaker's strategic interests.

I propose an environment within which agents can be evolved (i.e. trained) systematically in such a way that it facilitates the introspection of agent capabilities vis-à-vis disinformation and reasoning about adversary incentives to misrepresent their intentions. *Diplomacy*, a seven-player, incomplete information, perfect information board game, offers an ideal framework within which to study strategic reasoning and signalling via cheap talk. Recent work in reinforcement learning has demonstrated that autonomous agents can be evolved to play a variant of *Diplomacy* that omits player-to-player communication called No-Press *Diplomacy* [1], [2]. Building on this work by incrementally introducing the *Press*, player-to-player communication, would teach us how future autonomous agents will behave in an information environment filled with strategic actors and how autonomous agents themselves are able to strategically maneuver in the these environments.

I begin by proposing an ambitious one-year research plan for answering fundamental questions (Sections II-C and II-G) about how reinforcement learning agents reason about their own and their opponents' incentives for providing reliable information and disinformation. This is followed by a discussion, loosely framed within the formalism provided by game theory, of how *Diplomacy* differs from other environments (games) in which autonomous agents are trained.

II. Research Plan and Milestones

In Figure 1, I outline a proposed timeline for this effort along with achievable and substantively interesting milestones. Each enumerated component of the effort is described in more detail in a corresponding paragraph. The primary one-year effort presents a strategy from transitioning from No-Press *Diplomacy* to a (slightly limited) version of standard *Diplomacy* in such a way that each step allows for incremental understanding of how agents are adapting to shared private information and disinformation. Two extensions are also included, in less detail, to illustrate future efforts that will be enabled by this project.

A. No-Press Setup

The project begins by replicating previous work on reinforcement learning for No-Press *Diplomacy*. This allows us to establish a training and testing framework while ensuring we have matched the state-of-the-art for No-Press *Diplomacy* learning.

B. Best Policy Broadcast

In the subsequent step, the environment will be extended such that strategies can be broadcast publicly (i.e. to other agents). In evaluating this, the initial implementation will allow only the broadcast of previous best policies – in other words, bots cannot lie. Instead, agents will broadcast to opponents their previous best policy conditional on the current state of the board. This is not necessarily their current best move, but the best move that the previous iteration of the algorithm identifies under the current circumstances. Agent strategies are therefore conditional on opponent previous best

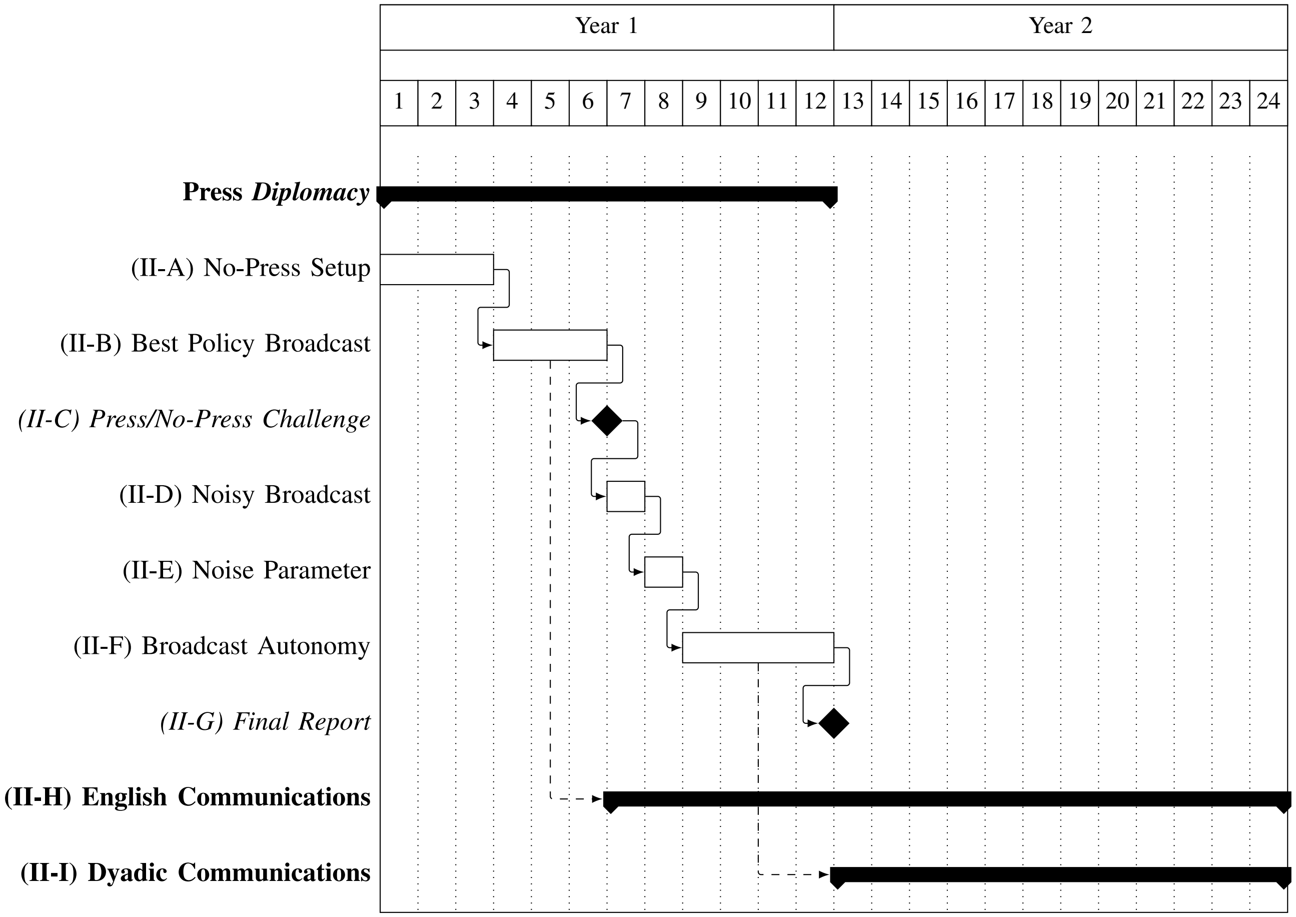

Fig. 1. Gantt chart depicting major research and engineering efforts and associated milestones. Detailed descriptions of each stage are referenced in parentheses.

strategies. Borrowing the notation of [1], we adapt the Q function:

$$Q_i^{\pi^b}(\alpha_i|s) \rightarrow Q_i^{\pi^b}(\alpha_i|s, \pi_{-i}^{t-1}) \tag{1}$$

by adding opponent previous best responses, π_{-i}^{t-1}, alongside s, the board state.

C. Press/No-Press Challenge

With the introduction of broadcast (public) communications, we can evaluate the ability of agents to adapt to the additional information about their opponents' strategies. A mid-project study will pit No-Press agents from step 1 against the press agents from step 2 and evaluate their performance against one another. A mid-project paper will answer questions including:

1) Do agents that broadcast their previous best policy (a) suffer due to other agents' ability to adapt or (b) benefit via signalling their "resolve" to follow through with their actions?
2) Do agents that do not broadcast their moves adjust their strategy set as a result of knowing some of their opponents' previous best policies?

D. Noisy Broadcast

In this stage, noise is added to best policy broadcasts. With probability $Pr(\text{lie}) = \alpha$, the agent broadcasts an erroneous previous best policy move selected from the set of valid moves; with probability $1 - \alpha$ the agent broadcasts the previous best policy move. Specifically, π_{-i}^{t-1} from Equation 1 is replaced with p_{-i}^{t-1} from Equation 2. $B(1, \alpha)$ denotes a single draw from a Bernoulli distribution with probability α. This does not allow agents to strategically select to use misinformation (the lie decision is exogenous), but it does allow us to study whether agents are able to detect false previous best policy moves.

$$p_{-i}^{t-1} = \begin{cases} \neg\pi_{-i}^{t-1} & B(1, \alpha) = 1 \\ \pi_{-i}^{t-1} & otherwise \end{cases} \tag{2}$$

For the remainder of the project, it is imperative that agents are able to weight the credibility of broadcasted move intentions better than chance. If agents are able to do so, it implies they are able to reason about their adversaries' incentives and strategies.

E. Noise Parameter

In this stage, probability that a true previous best policy is broadcasted is a learned parameter for each agent. Each agent can adjust their own α but, if a deceitful previous best policy is broadcasted, the agent has no control over the content of that broadcast – it is randomly chosen from the set of available strategies.

F. Broadcast Autonomy

Finally, the learned agent noise parameter α is replaced with the full set of available strategies. Agents therefore select two strategies for every iteration: the true strategy and the broadcasted strategy. These may or may not be the same. At this point, the game played by the agents is closer to standard *Diplomacy* than it is No-Press *Diplomacy*.

G. Final Report

A final paper details the successes and challenges of each previous above step. The focus of the paper is on resolving the following questions:

1) Does $\alpha = 0$ (i.e. no deceit) lead to a best second-place strategy as it allows opponent agents to adapt strategies to minimize expended effort countering the truthful bot? Alternatively, are agents able to leverage imperfect but good (previous best) strategy information about their opponents to eliminate them early in competition?
2) Do agents converge on a single α value? Is there an optimal amount of "good" information to broadcast?
3) Are agents able to select better-than-random deceitful moves? When agents in stage II-F select to broadcast a deceitful strategy, do they do so with greater success than agents in stage II-E?
4) Are agents able to detect broadcasted strategies that are deceitful (i.e. sub-optimal for the broadcaster)? In other words, are agents able to reason about their opponents' strategies? This can be shown via comparison of agent behaviors and success rates in stage II-F and stage II-F.

H. Natural Language Communications

Communications between agents in *Diplomacy* do not necessarily need to be made in natural language; in fact, the effort proposed above envisions only communications made using a structured vocabulary of legal moves. Reinforcement learning agents that operate in real-world information environments (e.g. social media) will need to translate (dis)information into natural human language. Once the *Press* is introduced in the *Diplomacy* laboratory environment in stage II-B, a parallel effort should explore bounding communications to human-like language. In other words: constraining agents to broadcast their intended (or, later, deceitful) strategies via language that is indistinguishable or nearly indistinguishable from the language used by human players. One approach to achieve this may be to train an auxiliary model to discriminate human-generated *Diplomacy* communications from agent-generated communications. This model may then be incorporated into the loss function used to update agent parameters. Each agent would then learn to play *Diplomacy* in parallel with a language generation model.

I. Dyadic Communications

Standard games of *Diplomacy* are complicated by the presence of shared private information. *Diplomacy* encourages players to share differing pieces of information with one another via dyadic communications. This facilitates the establishment of (temporary) alliances and makes possible explicit cooperation. It also allows players to strategically establish multilateral agreements upon which they later renege. However, this complicates the reinforcement learning problem because it requires a framework by which agents can communicate unique messages to one another. Each agent has up to 6 opponents for which it needs to convey seemingly consistent strategies. Furthermore, its possible for players to communicate not only their own strategies but also their opponents' privately-stated strategies to others. Maintaining a reinforcement learning environment that facilitates first- and, possibly, second-order dyadic communications poses a technical challenge beyond the scope of the initial project. However, doing so would allow for the study of much more complex agents that must consider not only how their opponents will interpret revealed strategies but how those opponents will communicate amongst themselves. Perhaps most interestingly, this would allow agents to explicitly cooperate in noisy cheap-talk environments and, possibly, develop their own reputation-based institutions to constrain their own strategic choices.

III. How is *Diplomacy* Different?

Reinforcement learning has shown great success in recent years with respect to developing autonomous agents capable of competing at or above the human expert level in a number of games. In this section, I provide preliminary thoughts on how *Press Diplomacy* differs from those games. While these thoughts are described using the formal language of game theory, proof of the game theoretic properties of *Diplomacy* and other games is left for a later research effort.

A. Poker

Poker and *Diplomacy* differ in substantial ways. Consider a game of *Diplomacy* versus a hand of no limit Texas Hold'em poker.[1] Game theory offers a typology of game characteristics that can be used to distinguish the two games. A few of these differences are elaborated here.

Most notably, *Diplomacy* is deterministic while Texas Hold'em is stochastic. *Diplomacy* starting positions are constant and player moves, along with the rule set, determine all subsequent board states. Texas Hold'em, on the other hand, incorporates a chance element: a shuffled card deck.

More useful for the purposes of evaluating conflict and cooperation in complex multi-agent information environments, poker offers players the ability to signal to one another using only costly and public mechanisms. Bids confer information to all players simultaneously and can't be retracted. Every bid incurs a cost (either an monetary value or an opportunity cost) and a single player cannot bid different amounts with respect

[1] The comparison works if extended to multiple hands of poker, too.

to different opponents. In other words, all communications in Poker are costly signals – they affect the game's final payoff.[2]

Diplomacy is designed to require private communications between players. Player A may signal strategy I to all players publicly and strategies II through VII to her opponents privately. Furthermore, these communications bear no direct impact on the game's payoffs (i.e. they are "cheap talk"). However, player-instituted arrangements may impose costs on communications; for example, players may require that they show one another their moves prior to submission in order to increase the costs associated with ensuring their own cooperation. *Diplomacy*, therefore, offers players a much richer strategy set vis-à-vis communications.

In order to ensure players' ability to engage in cheap talk, *Diplomacy* is a simultaneous game.[3] Players are unaware of each other's intended strategies when selecting their own strategies. Poker, on the other hand, is a sequential game: nature (the dealer) moves and then players taken turns betting.

Poker is an *imperfect, complete information* game. Players are unaware of one another's hands (imperfect information) but are aware of one another's utility functions, payoffs, and available strategies. *Diplomacy* is an *incomplete, perfect information* game. When selecting their strategies, all players are equally-well informed about the state of the game and all past states (i.e. *Diplomacy* is a perfect information game). However, players hold private information about themselves and about one another, making it an incomplete information game.

B. Go

Go is a two player competitive game. As such, there is not the opportunity for cooperation to emerge between opponents as there is in *Diplomacy*. Furthermore, informal communication plays little to no role in Go and reinforcement learning agents that play Go, like AlphaGo, do not incorporate an inter-player communication scheme beyond costly signalling via strategy selection (i.e. moves).

C. Starcraft II

Currently, the state of the art Starcraft II reinforcement learning agent, AlphaStar, plays only in one-on-one matches. Therefore, like AlphaGo, AlphaStar learns competitive play and is unable to cheaply communicate cooperative (or non-competitive) strategies with other players.

D. No Press Diplomacy

The No Press variant of *Diplomacy*, the version played by DeepMind's learning agents, omits from *Diplomacy* the entire negotiation phase of play. No Press forbids communications between players and therefore transforms the game in such a way that it no longer allows cheap talk. As with Poker, all signals in No Press *Diplomacy* are costly.

IV. Heilmeier Catechism

1) What are you trying to do? Articulate your objectives using absolutely no jargon.
   - We want a scalable solution for monitoring, moderating, (and/or influencing) online conversations in environments with intelligent adversaries. To do so, we will train autonomous agents ("bots") to consider their own and their opponent's information environments and strategies to (a) strategically convey believable misinformation and (b) detect truthful and deceptive information conditional on their opponents' (or allies') incentives and available strategies. In other words, we will train bots to lie compellingly and to detect likely falsehoods.
2) How is it done today, and what are the limits of current practice?
   - Bot and troll accounts on social media are either (a) unintelligent scripts or (b) closely managed by humans. Detection and remediation efforts are also largely manual. However, the low cost of managing bots (which requires less skill than detection and remediation), coupled with rapidly improving language generation technologies, means that bots engaged in disinformation campaigns will outpace the ability of moderators to effectively moderate fora.
3) What's new in your approach and why do you think it will be successful?
   - We will use reinforcement learning to train agents to play *Diplomacy*. While this has been accomplished previously, we extend the state of the art by training on the *Press* variant of *Diplomacy*. This change is substantively interesting because it allows us, for the first time, to study agents engaging in both costly and cheap talk via both public and private channels. Because the "language" of *Diplomacy* is formalized, we can determine which communications are accurate and which are inaccurate – we can therefore study the development of deceitful and cooperative strategies among agents. Furthermore, we anticipate being able to introspect agents' assessments of their opponents' signals; are agents able to distinguish credible signals from incredible signals?
4) Who cares? If you're successful, what difference will it make?
   - Autonomous agents already produce text that is nearly indistinguishable from text written by a human. However, the content of these generated texts is often not compelling and, over time, self-inconsistent. Given the incentives for states and other modestly-funded actors to gain control over social media and other information environments via high volume conversational bots, the quality of discourse these bots are capable of will likely

[2] This ignores cheating or body language, aspects not incorporated into reinforcement learning poker agents.

[3] It is a finitely repeated simultaneous game.

increase. A *Diplomacy* lab for studying strategic communication among autonomous agents allows for instrumentation at many points along one possible trajectory by which bots will evolve more advanced rhetorical capabilities (as outlined in Sections II-B through II-F). This will give researchers and policymakers an advantage with respect to both detecting the strategic use of disinformation by autonomous agents and developing training methods for producing convincing autonomous agents.

5) What are the risks and the payoffs?
    - TBD.
6) How much will it cost?
    - TBD.
7) How long will it take?
    - The proposed effort is anticipated to last one year. Additional follow-on work, described in II-H and II-I, is likely to take an additional year or longer.
8) What are the midterm and final ”exams” to check for success?
    - See Sections II-C and II-G.

References

[1] T. Anthony, T. Eccles, A. Tacchetti, J. Kramár, I. Gemp, T. C. Hudson, N. Porcel, M. Lanctot, J. Pérolat, R. Everett, S. Singh, T. Graepel, and Y. Bachrach, “Learning to play no-press diplomacy with best response policy iteration,” *arXiv:2006.04635v2*, 2020.

[2] P. Paquette, Y. Lu, S. Bocco, M. O. Smith, S. Ortiz-Gagné, J. K. Kummerfeld, S. Singh, J. Pineau, and A. Courville, “No press diplomacy: Modeling multi-agent gameplay,” *33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*, 2019.