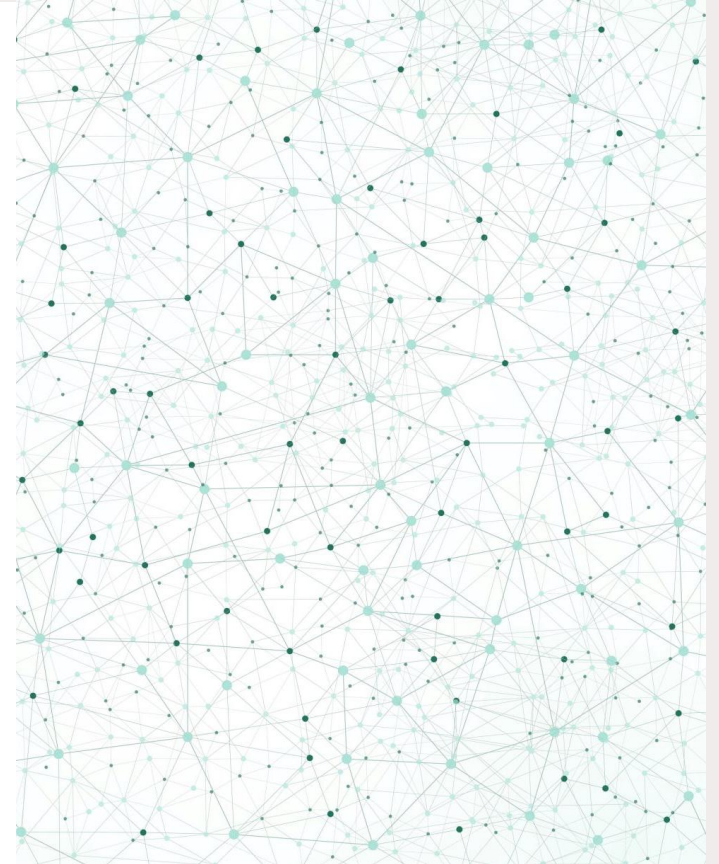


"Me lo dijo ChatGPT": sobre la capacidad de los modelos de lenguaje de afirmar

MAURO SANTELLI [(UBA - IIF-SADAF(CONICET))]



Estructura

- Introducción
- Loros estocásticos: el desafío de Emily Bender et al. a los grandes modelos de lenguaje
 - Algunas precisiones sobre lo que son los modelos de lenguaje actuales
- Teoría de la aserción y las condiciones de Bender
- La capacidad de afirmar
- Comentarios finales

Introducción

- ¿Qué queremos decir con que alguien nos dijo algo?
 - Esta pregunta es casi trivial cuando hablamos entre humanos
 - Con los modelos de lenguaje no tanto
- ¿Pueden usar el lenguaje los modelos de lenguaje?
 - Según varios argumentos existentes, no.
 - ¿pueden afirmar?
- ¿Qué es necesario para afirmar?
 - Principales propuestas sobre la aserción
 - ¿Cuáles nos sirven para dilucidar la capacidad de una máquina de afirmar?

Introducción

- Objetivos
 - No me propongo dilucidar de una vez y para siempre si los modelos de lenguaje son intrínsecamente capaces de afirmar.
 - Los modelos de lenguaje me van a servir para discutir las condiciones para la afirmación
 - Como corolario, mejores criterios sobre la afirmación podría servir para evaluar modelos de lenguaje.
 - Desarrollar evaluaciones estandarizadas, por ejemplo.

Con modelos de lenguaje me refiero a ChatGPT (2, 3, 4, 4o, o1, etc.), Gemini, Claude 3.5 opus, Llama 3, etc.

Loros estocásticos:



Loros estocásticos

On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?

Emily M. Bender*

ebender@uw.edu

University of Washington

Seattle, WA, USA

Timnit Gebru*

timnit@blackinai.org

Black in AI

Palo Alto, CA, USA

Angelina McMillan-Major

aymm@uw.edu

University of Washington

Seattle, WA, USA

Shmargaret Shmitchell

shmargaret.shmitchell@gmail.com

The Aether

Loros estocásticos

- La lingüista Emily Bender, junto a colegas, argumentó que los grandes modelos de lenguaje (hace tres años) suponían riesgos. Allí acuñó la analogía entre los grandes modelos de lenguaje (LLMs) y “loros estocásticos”.
- Con estocástico se refiere a la naturaleza probabilística de los modelos generativos.
 - Estos modelos se entrenan con una gran cantidad de datos.
 - Se fragmenta el texto en tokens
 - Se procesan los tokens en épocas (pasadas de procesamiento sobre los datos)
 - Se encuentran asociaciones probabilísticas entre los tokens
 - El modelo es capaz de actualizar sus predicciones del próximo token a partir de los anteriores. El último contiene la información actualizada de todos los tokens que lo antecedieron.

Loros estocásticos

- Estos modelos se entrenan con una gran cantidad de datos.
- Se fragmenta el texto en tokens (cosas como "papa" pueden ser un token, pero también "bol.", no necesariamente tienen que ser significativos para nosotros)
- Se procesan los tokens en épocas (pasadas de procesamiento sobre los datos)
- Se encuentran asociaciones probabilísticas entre los tokens
- El modelo es capaz de actualizar sus predicciones del próximo token a partir de los anteriores. El último contiene la información actualizada de todos los tokens que lo

Entrada:	La	53%	ponencia	40%	es	25%	Horrible
		40%	maravilla	30%	fue	25%	Clara
		5%	presidenta	30%	será	25%	Hermosa
		2%	Argentina			20%	Informativa
						10%	Improvisada
=	"La ponencia será horrible"						

Loros estocásticos

- Es conveniente leer el argumento de Bender et al. (2021) en algún detalle:

“Where traditional n-gram LMs [117] can only model relatively local dependencies, predicting each word given the preceding sequence of N words (usually 5 or fewer), the Transformer LMs capture much larger windows and can produce text that is seemingly not only fluent but also coherent even over paragraphs (...)

Loros estocásticos

- Es conveniente leer el argumento de Bender et al. (2021) en algún detalle:

“Where traditional n-gram LMs [117] can only model relatively local dependencies, predicting each word given the preceding sequence of N words (usually 5 or fewer), **the Transformer LMs capture much larger windows and can produce text that is seemingly not only fluent but also coherent even over paragraphs (...)**

Loros estocásticos

- Es conveniente leer el argumento de Bender et al. (2021) en algún detalle:
- We say seemingly coherent because coherence is in fact in the eye of the beholder. Our human understanding of coherence derives from our ability to recognize interlocutors' beliefs [30, 31] and intentions [23, 33] within context [32]. That is, **human language use takes place between individuals who share common ground and are mutually aware of that sharing (and its extent), who have communicative intents which they use language to convey, and who model each others' mental states as they communicate**. As such, **human communication relies on the interpretation of implicit meaning conveyed between individuals**. The fact that human-human communication is a jointly constructed activity [29, 128] is most clearly true in co-situated spoken or signed communication, but we use the same facilities for producing language that is intended for audiences not co-present with us (readers, listeners, watchers at a distance in time or space) and in interpreting such language when we encounter it. It must follow that even when we don't know the person who generated the language we are interpreting, we build a partial model of who they are and what common ground we think they share with us, and use this in interpreting their words." (p. 616)

Loros estocásticos

- Es conveniente leer el argumento de Bender et al. (2021) en algún detalle:
- **Text generated by an LM is not grounded in communicative intent, any model of the world, or any model of the reader's state of mind.** It can't have been, because the training data never included sharing thoughts with a listener, nor does the machine have the ability to do that. This can seem counter-intuitive given the increasingly fluent qualities of automatically generated text, but we have to account for the fact that our perception of natural language text, regardless of how it was generated, is mediated by our own linguistic competence and our predisposition to interpret communicative acts as conveying coherent meaning and intent, whether or not they do [89, 140]. The problem is, if one side of the communication does not have meaning, then the comprehension of the implicit meaning is an illusion arising from our singular human understanding of language

Loros estocásticos

- Es conveniente leer el argumento de Bender et al. (2021) en algún detalle:
- Contrary to how it may seem when we observe its output, an LM is a system for haphazardly stitching together sequences of linguistic forms it has observed in its vast training data, according to probabilistic information about how they combine, but without any reference to meaning: a stochastic parrot. (617)

Loros estocásticos

- Es conveniente leer el argumento de Bender et al. (2021) en algún detalle:
- Contrary to how it may seem when we observe its output, an LM is a system for haphazardly stitching together sequences of linguistic forms it has observed in its vast training data, according to probabilistic information about how they combine, but without any reference to meaning: **a stochastic parrot.** (617)
- “We note that the risks associated with synthetic but seemingly coherent text are deeply connected to the fact that **such synthetic text can enter into conversations without any person or entity being accountable for it.** This accountability both involves responsibility for truthfulness and is important in situating meaning. As Maggie Nelson [92] writes: “Words change depending on who speaks them; there is no cure.”

Loros estocásticos

- Es conveniente leer el argumento de Bender et al. (2021) en algún detalle:
- Contrary to how it may seem when we observe its output, an LM is a system for haphazardly stitching together sequences of linguistic forms it has observed in its vast training data, according to probabilistic information about how they combine, but without any reference to meaning: **a stochastic parrot.** (617)
- “We note that the risks associated with synthetic but seemingly coherent text are deeply connected to the fact that such synthetic text can enter into conversations without any person or entity being accountable for it. **This accountability both involves responsibility for truthfulness and is important in situating meaning.** As Maggie Nelson [92] writes: “Words change depending on who speaks them; there is no cure.”

Loros estocásticos

- Es conveniente leer el argumento de Bender et al. (2021) en algún detalle:
- Contrary to how it may seem when we observe its output, an LM is a system for haphazardly stitching together sequences of linguistic forms it has observed in its vast training data, according to probabilistic information about how they combine, but without any reference to meaning: **a stochastic parrot.** (617)
- “We note that the risks associated with synthetic but seemingly coherent text are deeply connected to the fact that such synthetic text can enter into conversations without any person or entity being accountable for it. **This accountability both involves responsibility for truthfulness and is important in situating meaning.** As Maggie Nelson [92] writes: “Words change depending on who speaks them; there is no cure.”

Loros estocásticos

- Es conveniente leer el argumento de Bender et al. (2021) en algún detalle:
- We say seemingly coherent because coherence is in fact in the eye of the beholder. Our human understanding of coherence derives from our ability to recognize interlocutors' beliefs [30, 31] and intentions [23, 33] within context [32]. That is, **human language use takes place between individuals who share common ground and are mutually aware of that sharing** (and its extent), who have communicative intents which they use language to convey, and who model each others' mental states as they communicate. As such, **human communication relies on the interpretation of implicit meaning conveyed between individuals**. The fact that human-human communication is a jointly constructed activity [29, 128] is most clearly true in co-situated spoken or signed communication, but we use the same facilities for producing language that is intended for audiences not co-present with us (readers, listeners, watchers at a distance in time or space) and in interpreting such language when we encounter it. It must follow that even when we don't know the person who generated the language we are interpreting, we build a partial model of who they are and what common ground we think they share with us, and use this in interpreting their words." (p. 616)

Loros estocásticos

- Es conveniente leer el argumento de Bender et al. (2021) en algún detalle:
- We say seemingly coherent because coherence is in fact in the eye of the beholder. Our human understanding of coherence derives from our ability to recognize interlocutors' beliefs [30, 31] and intentions [23, 33] within context [32]. That is, 1) **human language use takes place between individuals who share common ground** and are mutually aware of that sharing (and its extent), who have communicative intents which they use language to convey, and who model each others' mental states as they communicate. As such, **human communication relies on the interpretation of implicit meaning conveyed between individuals**. The fact that human-human communication is a jointly constructed activity [29, 128] is most clearly true in co-situated spoken or signed communication, but we use the same facilities for producing language that is intended for audiences not co-present with us (readers, listeners, watchers at a distance in time or space) and in interpreting such language when we encounter it. It must follow that even when we don't know the person who generated the language we are interpreting, we build a partial model of who they are and what common ground we think they share with us, and use this in interpreting their words." (p. 616)

Loros estocásticos

- 1) Comparten un trasfondo común
 - a. Y están conscientes de ese hecho.

Loros estocásticos

- Es conveniente leer el argumento de Bender et al. (2021) en algún detalle:
- We say seemingly coherent because coherence is in fact in the eye of the beholder. Our human understanding of coherence derives from our ability to recognize interlocutors' beliefs [30, 31] and intentions [23, 33] within context [32]. That is, 1) **human language use takes place between individuals who share common ground** and are mutually aware of that sharing (and its extent), who 2) have communicative intents which they use language to convey, and who model each others' mental states as they communicate. As such, **human communication relies on the interpretation of implicit meaning conveyed between individuals**. The fact that human-human communication is a jointly constructed activity [29, 128] is most clearly true in co-situated spoken or signed communication, but we use the same facilities for producing language that is intended for audiences not co-present with us (readers, listeners, watchers at a distance in time or space) and in interpreting such language when we encounter it. It must follow that even when we don't know the person who generated the language we are interpreting, we build a partial model of who they are and what common ground we think they share with us, and use this in interpreting their words." (p. 616)

Loros estocásticos

- 1) Comparten un trasfondo común
 - a. Y están conscientes de ese hecho.
- 2) Tienen intenciones comunicativas
 - a. Tienen *algo* que decir por *medio* del lenguaje

Loros estocásticos

- Es conveniente leer el argumento de Bender et al. (2021) en algún detalle:
- We say seemingly coherent because coherence is in fact in the eye of the beholder. Our human understanding of coherence derives from our ability to recognize interlocutors' beliefs [30, 31] and intentions [23, 33] within context [32]. That is, 1) **human language use takes place between individuals who share common ground** and are mutually aware of that sharing (and its extent), who 2) **have communicative intents which they use language to convey**, and 3) **who model each others' mental states as they communicate**. As such, **human communication relies on the interpretation of implicit meaning conveyed between individuals**. The fact that human-human communication is a jointly constructed activity [29, 128] is most clearly true in co-situated spoken or signed communication, but we use the same facilities for producing language that is intended for audiences not co-present with us (readers, listeners, watchers at a distance in time or space) and in interpreting such language when we encounter it. It must follow that even when we don't know the person who generated the language we are interpreting, we build a partial model of who they are and what common ground we think they share with us, and use this in interpreting their words." (p. 616)

Loros estocásticos

- 1) Comparten un trasfondo común
 - a. Y están conscientes de ese hecho.
- 2) Tienen intenciones comunicativas
 - a. Tienen *algo* que decir por *medio* del lenguaje
- 3) Tienen un modelo mental de los estados doxásticos del otro

Loros estocásticos

- 1) Comparten un trasfondo común
 - a. Y están conscientes de ese hecho.
- 2) Tienen intenciones comunicativas
 - a. Tienen *algo* que decir por *medio* del lenguaje
- 3) Tienen un modelo mental de los estados doxásticos del otro

Esto redundaría en que la comunicación humana descansa en la interpretación de **significado implícito** transmitido entre los individuos.

Loros estocásticos

“We note that the risks associated with synthetic but seemingly coherent text are deeply connected to the fact that **such synthetic text can enter into conversations without any person or entity being accountable for it**. This accountability both involves responsibility for truthfulness and is important in situating meaning. As Maggie Nelson [92] writes: “Words change depending on who speaks them; there is no cure.”

Loros estocásticos

- 1) Comparten un trasfondo común
 - a. Y están conscientes de ese hecho.
- 2) Tienen intenciones comunicativas
 - a. Tienen *algo* que decir por *medio* del lenguaje
- 3) Tienen un modelo mental de los estados doxásticos del otro

Esto redundo en que la comunicación humana descansa en la interpretación de **significado implícito** transmitido entre los individuos.

- 4) Responsabilidad
 - a. Fáctica
 - b. Ubicación del significado. Determinación del significado por autoría. “Las palabras cambian dependiendo de quién las dice; no hay cura”.

Cuatro modelos de la aserción

John MacFarlane (2010) presenta cuatro grandes modelos de teorías de la aserción:

1. Modelo actitudinal: Afirmar es expresar una actitud
2. Modelo de reglas constitutivas: afirmar es hacer una jugada definida por sus reglas constitutivas
3. Modelo del contexto común: Afirmar es proponer agregar información al contexto conversacional común.
4. Modelo del compromiso: Afirmar es asumir un compromiso

Cuatro modelos de la aserción

John MacFarlane (2010) presenta cuatro grandes modelos de teorías de la aserción:

1. Modelo actitudinal: Afirmar es expresar una actitud (Bach y Harnish, B. Williams, etc.)
2. Modelo de reglas constitutivas: afirmar es hacer una jugada definida por sus reglas constitutivas (T. Williamson, Lackey, Price, etc.)
3. Modelo del contexto común: Afirmar es proponer agregar información al contexto conversacional común. (Stalnaker, ¿Lewis?)
4. Modelo del compromiso: Afirmar es asumir un compromiso (Peirce, Searle, Brandom, entre otros).

Cuatro **Cinco** modelos de la aserción

John MacFarlane (2011) presenta cuatro grandes modelos de teorías de la aserción. Jessica Brown y Herman Cappelen (2011) establecen algunas diferencias.

1. Modelo actitudinal / **Caracterización por las causas** (Bach y Harnish, B. Williams, etc.)
2. Modelo de reglas constitutivas / **Caracterización normativa**: afirmar es hacer una jugada definida por sus reglas constitutivas (T. Williamson, Lackey, Price, etc.)
3. Modelo del contexto común / **Caracterización por los efectos**: Afirmar es proponer agregar información al contexto conversacional común. (Stalnaker, ¿Lewis?)
4. Modelo del compromiso: Afirmar es asumir un compromiso (Peirce, Searle, Brandom, entre otros).
5. Inexistencia de la aserción: La aserción no es un fenómeno real.

~~Cuatro~~ **Seis (y medio)** modelos de la aserción

John MacFarlane (2010) presenta cuatro grandes modelos de teorías de la aserción. Jessica Brown y Herman Cappelen (2010) establecen algunas diferencias. **A su vez, propuestas recientes sugieren algunos cambios.**

1. Modelo actitudinal / Caracterización por las causas (Bach y Harnish, B. Williams, etc.)
2. Modelo de reglas constitutivas / Caracterización normativa: afirmar es hacer una jugada definida por sus reglas constitutivas (T. Williamson, Lackey, Price, etc.)
 - a. Modelos normativos en donde la norma es una función que va de contextos a normas inesenciales pero contextualmente relevantes. (Janet Levin, Pagin)
3. Modelo del contexto común / Caracterización por los efectos: Afirmar es proponer agregar información al contexto conversacional común. (Stalnaker, ¿Lewis?)
4. Modelo del compromiso: Afirmar es asumir un compromiso (Peirce, Searle, Brandom, entre otros).
5. Inexistencia de la aserción: La aserción no es un fenómeno real.
- 6. Modelos híbridos: la aserción es un fenómeno complejo que involucra todos los elementos de 1-4.** (María José Frápolli (2023) defiende este modelo en términos de Frege y de Brandom).

~~Cuatro~~ **Seis (y medio)** modelos de la aserción

John MacFarlane (2010) presenta cuatro grandes modelos de teorías de la aserción. Jessica Brown y Herman Cappelen (2010) establecen algunas diferencias. **A su vez, propuestas recientes sugieren algunos cambios.**

1. ~~Modelo actitudinal / Caracterización por las causas (Bach y Harnish, B. Williams, etc.)~~
2. Modelo de reglas constitutivas / **Caracterización normativa:** afirmar es hacer una jugada definida por sus reglas constitutivas (T. Williamson, Lackey, Price, etc.)
 - a. Modelos normativos en donde la norma es una función que va de contextos a normas inesenciales pero contextualmente relevantes. (Janet Levin, Pagin)**
3. Modelo del contexto común / **Caracterización por los efectos:** Afirmar es proponer agregar información al contexto conversacional común. (Stalnaker, ¿Lewis?)
4. Modelo del compromiso: Afirmar es asumir un compromiso (Peirce, Searle, Brandom, entre otros).
5. ~~Inexistencia de la aserción: La aserción no es un fenómeno real.~~
- 6. Modelos híbridos: la aserción es un fenómeno complejo que involucra todos los elementos de 1-4. (María José Frápolli (2023) defiende este modelo en términos de Frege y de Brandom).**

~~Cuatro~~ **Seis (y medio)** modelos de la aserción

John MacFarlane (2010) presenta cuatro grandes modelos de teorías de la aserción. Jessica Brown y Herman Cappelen (2010) establecen algunas diferencias. **A su vez, propuestas recientes sugieren algunos cambios.**

1. ~~Modelo actitudinal / Caracterización por las causas (Bach y Harnish, B. Williams, etc.)~~
2. ¿Modelo de reglas constitutivas / **Caracterización normativa:** afirmar es hacer una jugada definida por sus reglas constitutivas (T. Williamson, Lackey, Price, etc.)?
 a. **Modelos normativos en donde la norma es una función que va de contextos a normas inesenciales pero contextualmente relevantes. (Janet Levin, Pagin)**
3. Modelo del contexto común / **Caracterización por los efectos:** Afirmar es proponer agregar información al contexto conversacional común. (Stalnaker, ¿Lewis?)
4. Modelo del compromiso: Afirmar es asumir un compromiso (Peirce, Searle, Brandom, entre otros).
5. ~~Inexistencia de la aserción: La aserción no es un fenómeno real.~~
6. **Modelos híbridos: la aserción es un fenómeno complejo que involucra todos los elementos de 1-4.** (María José Frápolli (2023) defiende este modelo en términos de Frege y de Brandom).

~~Cuatro~~ **Seis (y medio)** modelos de la aserción

John MacFarlane (2010) presenta cuatro grandes modelos de teorías de la aserción. Jessica Brown y Herman Cappelen (2010) establecen algunas diferencias. **A su vez, propuestas recientes sugieren algunos cambios.**

1. ~~Modelo actitudinal / Caracterización por las causas (Bach y Harnish, B. Williams, etc.)~~
2. ¿Modelo de reglas constitutivas / **Caracterización normativa:** afirmar es hacer una jugada definida por sus reglas constitutivas (T. Williamson, Lackey, Price, etc.)?
 a. **Modelos normativos en donde la norma es una función que va de contextos a normas inesenciales pero contextualmente relevantes. (Janet Levin, Pagin)**
3. **Modelo del contexto común / Caracterización por los efectos: Afiramar es proponer agregar información al contexto conversacional común. (Stalnaker, ¿Lewis?)**
4. Modelo del compromiso: Afiramar es asumir un compromiso (Peirce, Searle, Brandom, entre otros).
5. ~~Inexistencia de la aserción: La aserción no es un fenómeno real.~~
6. **Modelos híbridos: la aserción es un fenómeno complejo que involucra todos los elementos de 1-4.** (María José Frápolli (2023) defiende este modelo en términos de Frege y de Brandom).

Cuatro Seis (y medio) modelos de la aserción

John MacFarlane (2010) presenta cuatro grandes modelos de teorías de la aserción. Jessica Brown y Herman Cappelen (2010) establecen algunas diferencias. **A su vez, propuestas recientes sugieren algunos cambios.**

1. ~~Modelo actitudinal / Caracterización por las causas (Bach y Harnish, B. Williams, etc.)~~
2. ¿Modelo de reglas constitutivas / Caracterización normativa: afirmar es hacer una jugada definida por sus reglas constitutivas (T. Williamson, Lackey, Price, etc.)?
 a. Modelos normativos en donde la norma es una función que va de contextos a normas inesenciales pero contextualmente relevantes. (Janet Levin, Pagin)
3. Modelo del contexto común / Caracterización por los efectos: Afirmar es proponer agregar información al contexto conversacional común. (Stalnaker, ¿Lewis?)
4. Modelo del compromiso: Afirmar es asumir un compromiso (Peirce, Searl)
5. ~~Inexistencia de la aserción: La aserción no es un fenómeno real.~~
6. Modelos híbridos: la aserción es un fenómeno complejo que involucra todo este modelo en términos de Frege y de Brandom).

Parece faltar la dimensión de la responsabilidad personal. Pero es compatible.

Cuatro **Seis (y medio)** modelos de la aserción

John MacFarlane (2010) presenta cuatro grandes modelos de teorías de la aserción. Jessica Brown y Herman Cappelen (2010) establecen algunas diferencias. **A su vez, propuestas recientes sugieren algunos cambios.**

1. Modelo actitudinal / ~~Caracterización por las causas~~ (Bach y Harnish, B. Williams, etc.)
2. ¿Modelo de reglas constitutivas / **Caracterización normativa**: afirmar es hacer una jugada definida por sus reglas constitutivas (T. Williamson, Lackey, Price, etc.)?
 a. **Modelos normativos en donde la norma es una función que va de contextos a normas inesenciales pero contextualmente relevantes.** (Janet Levin, Pagin)
3. Modelo del contexto común / **Caracterización por los efectos**: Afirmar es proponer agregar información al contexto conversacional común. (Stalnaker, ¿Lewis?)
4. Modelo del compromiso: Afirmar es asumir un compromiso (Peirce, Searle, Brandom, entre otros).
5. ~~Inexistencia de la aserción~~: La aserción no es un fenómeno real.
6. **Modelos híbridos: la aserción es un fenómeno complejo que involucra todos los elementos de 1-4.** (María José Frápolli (2023) defiende este modelo en términos de Frege y de Brandom).

Loros estocásticos

- 1) Comparten un trasfondo común
 - a. Y están conscientes de ese hecho.
- 2) Tienen intenciones comunicativas
 - a. Tienen *algo* que decir por *medio* del lenguaje
- 3) Tienen un modelo mental de los estados doxásticos del otro

Esto redundaría en que la comunicación humana descansa en la interpretación de **significado implícito** transmitido entre los individuos.

- 4) Responsabilidad
 - a. Fáctica
 - b. Ubicación del significado. Determinación del significado por autoría. “Las palabras cambian dependiendo de quién las dice; no hay cura”.

¿Cuáles son exigibles a un modelo de lenguaje?

- 1) Comparten un trasfondo común
 - a. Y están conscientes de ese hecho.
- 2) Tienen intenciones comunicativas
 - a. Tienen *algo* que decir por *medio* del lenguaje
- 3) Tienen un modelo mental de los estados doxásticos del otro

Esto redundaría en que la comunicación humana descansa en la interpretación de **significado implícito** transmitido entre los individuos.

- 4) Responsabilidad
 - a. Fáctica
 - b. Ubicación del significado. Determinación del significado por autoría. “Las palabras cambian dependiendo de quién las dice; no hay cura”.

¿Cuáles son exigibles a un modelo de lenguaje?

- 1) Comparten un trasfondo común
 - a. Y están conscientes de ese hecho.
- 2) **¿Tienen intenciones comunicativas?**
 - a. Tienen *algo* que decir por *medio* del lenguaje
- 3) Tienen un modelo mental de los estados doxásticos del otro

Esto redundaría en que la comunicación humana descansa en un **implícito** transmitido entre los individuos.

- 4) Responsabilidad
 - a. Fáctica
 - b. Ubicación del significado. Determinación del significado por autoría. “Las palabras cambian dependiendo de quién las dice; no hay cura”.

¿Cómo podríamos determinar que un sujeto tiene intenciones comunicativas? Especialmente uno que no tiene cuerpo.

¿Cómo caracterizar ese algo que tienen para decir independientemente de su evaluación por medio de lo que dicen?

¿Cuáles son exigibles a un modelo de lenguaje?

Requisito de las intenciones comunicativas:

Modelar al otro parece ser un rasgo identificado por varias teorías de la aserción.

Pero el sub-requisito de tener *algo* que decir requiere una especificación en términos de creencias o ítems semánticos no lingüísticos que expresar.

Esto parece demasiado. Más bien creo que hay que adoptar una concepción contrapositiva:

Si decimos que ChatGPT no *afirma*, decimos que *no tiene intenciones comunicativas*.

¿Cuáles son exigibles a un modelo de lenguaje?

Si tiene intenciones comunicativas (entre otras condiciones), ChatGPT afirma [o usa el lenguaje]

Formulación contrapositiva del condicional pragmático:

"Si decimos que ChatGPT no afirma, decimos que no tiene intenciones comunicativas."

¿Cuáles son exigibles a un modelo de lenguaje?

- 1) Comparten un trasfondo común
 - a. Y están conscientes de ese hecho.
- 2) ***Si ChatGPT no afirma, no tiene intenciones comunicativas.**
 - a. Tienen *algo* que decir por *medio* del lenguaje
- 3) Tienen un modelo mental de los estados doxásticos del otro

Esto redunda en que la comunicación humana descansa en la interpretación de **significado implícito** transmitido entre los individuos.

- 4) Responsabilidad
 - a. Fáctica
 - b. Ubicación del significado. Determinación del significado por autoría. “Las palabras cambian dependiendo de quién las dice; no hay cura”.

¿Cuáles son exigibles a un modelo de lenguaje?

- 1) Comparten un trasfondo común
 - a. Y están conscientes de ese hecho.
- 2) ***Si ChatGPT no afirma, no tiene intenciones comunicativas.**
 - a. Lo que sea que le atribuyamos a ChatGPT va a estar en función de nuestra capacidad de trato a ChatGPT.
 - a. No es independiente de punto 1, 3 y 4.
- 3) Tienen un modelo mental de los estados doxásticos del otro

Esto redundaría en que la comunicación humana descansa en la interpretación de **significado implícito** transmitido entre los individuos.

- 4) Responsabilidad
 - a. Fáctica
 - b. Ubicación del significado. Determinación del significado por autoría. “Las palabras cambian dependiendo de quién las dice; no hay cura”.

¿Cuáles son exigibles a un modelo de lenguaje?

- 1) Comparten un trasfondo común
 - a. Y están conscientes de ese hecho.
- 2) ***Si ChatGPT no afirma, no tiene intenciones comunicativas.**
 - a. Lo que sea que le atribuyamos a ChatGPT va a estar en función de lo que sea que le atribuyamos a ChatGPT.
 - a. No es independiente de punto 1, 3 y 4.
- 3) **Tienen un modelo mental de los estados doxásticos del otro.**

Esto no sólo es algo que hacen ya, sino que es demandable.

Esto redunda en que la comunicación humana descansa en la interpretación de **significado implícito** transmitido entre los individuos.

4) Responsabilidad

- a. Fáctica
- b. Ubicación del significado. Determinación del significado por autoría. “Las palabras cambian dependiendo de quién las dice; no hay cura”.

¿Cuáles son exigibles a un modelo de lenguaje?

- 1) **Comparten un trasfondo común**
 - a. Y están conscientes de ese hecho.
- 2) ***Si ChatGPT no afirma, no tiene intenciones comunicativas.**
 - a. Lo que sea que le atribuyamos a ChatGPT va a estar en función de nuestra capacidad de trato a ChatGPT.
 - a. No es independiente de punto 1, 3 y 4.
- 3) **Tienen un modelo mental de los estados doxásticos del otro.**

Esto redunda en que la comunicación humana descansa en la interpretación de **significado implícito** transmitido entre los individuos.

- 4) Responsabilidad
 - a. Fáctica
 - b. Ubicación del significado. Determinación del significado por autoría. “Las palabras cambian dependiendo de quién las dice; no hay cura” (Maggie Nelson).

¿Cuáles son exigibles a un modelo de lenguaje?

1) Comparten un trasfondo común

- a. Y están conscientes de ese hecho.

2) ***Si ChatGPT no afirma, no tiene intenciones comunicativas.**

- a. Lo que sea que le atribuyamos a ChatGPT va a estar en función de nuestra capacidad de trato a ChatGPT.
 - a. No es independiente de punto 1, 3 y 4.

3) Tienen un modelo mental de los estados doxásticos del otro.

Esto redundo en que la comunicación humana descansa en la interpretación de **significado implícito** transmitido entre los individuos.

4) Responsabilidad

- a. Fáctica
- b. **Ubicación del significado. Determinación del significado por autoría. “Las palabras cambian dependiendo de quién las dice; no hay cura” (Maggie Nelson).**

¿Qué tipo de accountability es exigible a un modelo de lenguaje?



¿Qué tipo de accountability es exigible a un modelo de lenguaje?



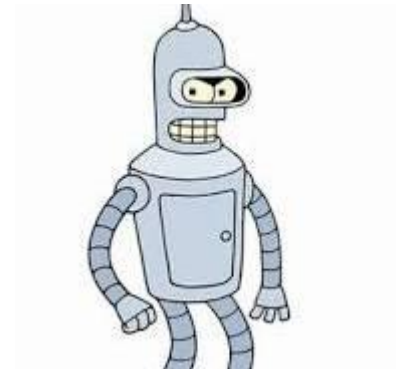
¿Cuáles son exigibles a un modelo de lenguaje?

- 1) **Comparten un trasfondo común**
 - a. Y están conscientes de ese hecho.
- 2) ***Si ChatGPT no afirma, no tiene intenciones comunicativas.**
 - a. Lo que sea que le atribuyamos a ChatGPT va a estar en función de nuestra capacidad de trato a ChatGPT.
 - a. No es independiente de punto 1, 3 y 4.
- 3) **Tienen un modelo mental de los estados doxásticos del otro.**

Esto redundaría en que la comunicación humana descansa en la interpretación de **significado implícito** transmitido entre los individuos.

- 4) Responsabilidad
 - a. Fáctica
 - b. **Ubicación del significado. Determinación del significado por autoría. “Las palabras cambian dependiendo de quién las dice; no hay cura” (Maggie Nelson).**
 - c. **Se requiere que los modelos defiendan sus posiciones afirmadas.**

La demanda de Bender



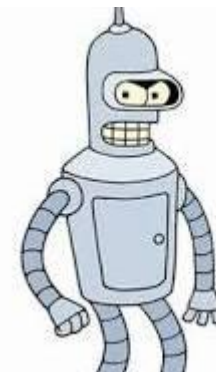
Bender parece exigirle a ChatGPT ser un agente, por lo que si la acción lingüística es condición suficiente para la agencia podemos preguntarnos directamente por la capacidad de aseverar.

Si fuera así, ¿qué le falta a un modelo de lenguaje?

Parece ser que lo requerido es cierta constancia en las creencias atribuibles a los agentes, para ser agentes.

Un modelo de lenguaje puede asumir diferentes personajes pero ninguno en serio.

Interrogantes abiertos



¿Es razonable pedirle identidad personal a un modelo del lenguaje?

¿Cómo se podría evaluar la capacidad de afirmación de un modelo del lenguaje?

¿Qué sería un examen de inserción?

Task 4: Two Argument Relations

The office is north of the bedroom.

The bedroom is north of the bathroom.

The kitchen is west of the garden.

What is north of the bedroom? A: office

What is the bedroom north of? A: bathroom

Task 6: Yes/No Questions

John moved to the playground.

Daniel went to the bathroom.

John went back to the hallway.

Is John in the playground? A:no

Is Daniel in the bathroom? A:yes

¡Muchas gracias!



Referencias

- Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? 🦜 .
In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency* (pp. 610-623).
- Brandom, R. (1998 [1994]). *Making it Explicit*. Harvard University Press
- Brown, J., & Cappelen, H. (Eds.). (2011). *Assertion: New philosophical essays*. Oxford University Press.
- Frápolli, M. J. (2023). *The Priority of Propositions. A Pragmatist Philosophy of Logic* (Vol. 475). Springer Nature.