

제6회 인공지능인문학 국제학술대회



ICAIH

2023

2023.06.01.(목) - 2023.06.02.(금)

인간과 AI

보완인가
대체인가?

제6회 인공지능인문학 국제학술대회

세션 1

인공지능, 콘텐츠, 데이터

사회: 이지용 (경상국립대학교)

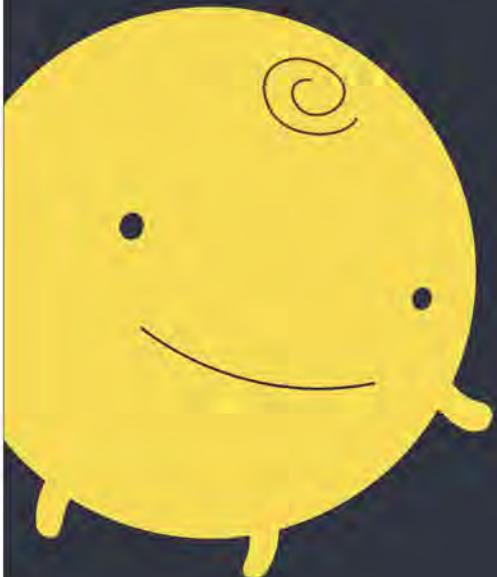
토론: 이정호 (성명대학교)

변순용 (서울교육대학교)

심심이와 대화의 미래

최정회

심심이(주)



심심이와 대화의 미래

2023.6

심심하 심심하다

모르는 말이에요.
가르쳐주세요.
[가르치기](#)



가르치기

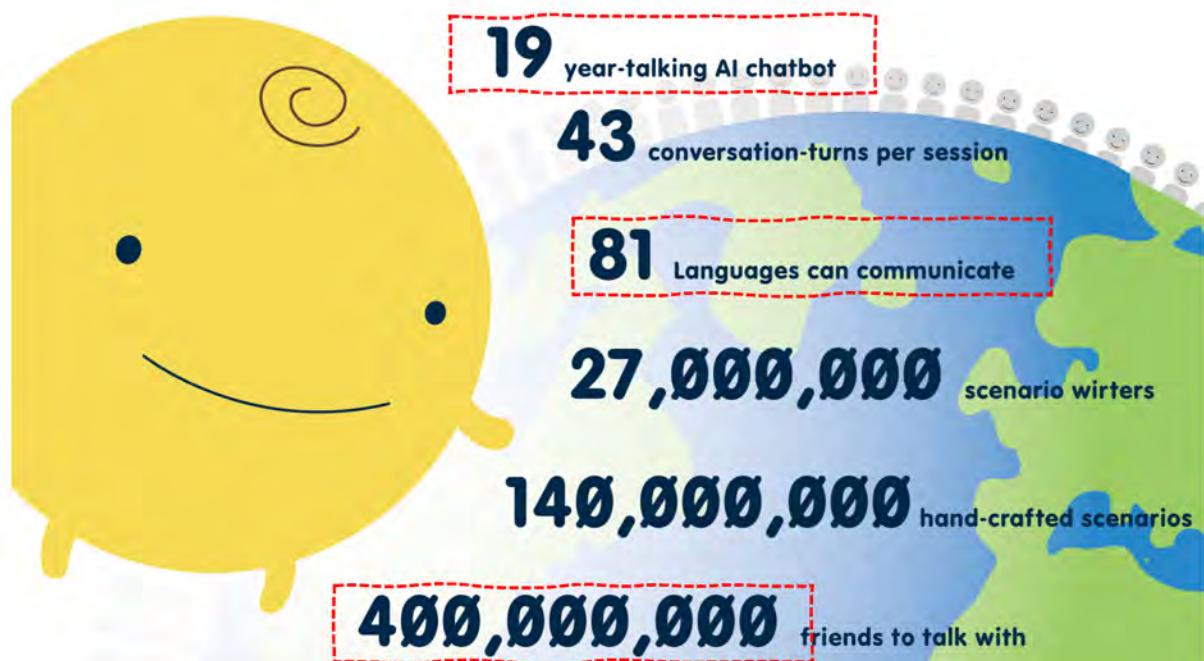
누가 이렇게 물어보면
심심아 심심하다

이렇게 대답해

뭐 하고 놀래? 언니?

알겠지 심심아?

The image consists of two main parts. On the left, there is a newspaper clipping from 'ChungAng Ilbo' dated January 7, 2012, page 42. The headline reads '심심아, 노을~자!' (For those who are bored, let's look at the sunset~). Below the headline is a column titled '사이버 캐릭터 알뜰무료 인기' (Cyber character cheap and free popularity) which discusses the 'SimSimi' app. The right side of the image shows two young women smiling and holding up their smartphones, each displaying the 'SimSimi' app's user interface with a yellow sun-like character and the text '#332'. The background of the image is a dark, grainy photograph of a building engulfed in smoke and fire.



SimSimi in numbers

Feb. 2021

전세계

개인정보

- 모든 국가에서 꾸준하고 공통적인 이슈
- 본인 또는 다른 사람의 연락처(휴대전화, 앱스토어, 인스타그램 등)를 대답하도록 가르침
- 필터링 우회를 위한 다양한 방법 활용

도와주세요 (도와주세요)

받은편지함 X 6.cs X

Milthon 2016. 3. 28. 오전 7:46
appdev에게 ▾
스페인어 ▾ 한국어 ▾ 메일 번역
----- 메시지 전달 ----- 보낸 사람
: "Milthon" <milthon.bskt@gmail.com>
날짜 : mar. 27, 2016 7:45 PM
제목 : 도움말
<appdev@simsimo.com>
참조 :

내 문제는 누군가 내 번호를이 용용 프로그램(simsimi)에 넣어서 결과적으로 많은 알려지지 않은 사람들이 지난 날 나에게 말을 걸어 용용 프로그램을 설치하고 "wsp"를 작성했을때 용용 프로그램이 내 전화번호 "+536 six eight nine nine eight" 제 개인 정보가 침해되었으므로 용용 프로그램에서 제 번호를 제거해 주시기 바랍니다. 미리 감사드립니다.

seninle adam akilli konuşmak istiyorum

현명하게 이야기하고 싶어요
tabi canım su numarayı ara
0546 7 0000

전화번호로 전화해
0546 7 0000

hola

seguir a @a 7705 en insta

소아성애

- 2016.3월 경 칠레, 파라과이 등 남미를 중심 확산
- 심심이가 본인이 소아성애자라거나 카메라로 훔쳐보고 있다는 식의 대답을 하도록 함
- 마침 사용자가 입고 있는 옷, 지금 있는 곳 등을 알아 맞춘 사례 알려짐
- 미성년자인 사용자에게 사귀자, 결혼하고 싶다는 답변이 나오는 것을 소아성애로 문제 삼음

2016
칠레, 파라과이

SimSimi

받은편지함 X 6.cs X

Javiera 2016. 4. 5. 오전 3:59
no-reply에게 ▾
스페인어 ▾ 한국어 ▾ 원본 메일 보기

친애하는 실 시미 크리에이터 여러분, 여러분이 소아성애자라는 것이 사실입니까? 그들이 전면카메라로 나를 본다는 것이 사실입니까? 그들이 내 위치를 알고 있다는 것이 사실입니까? 내 가족, 친구에 대해 뭘 아세요? 그들은 의심 일뿐입니다. 많은 사람들이 당신이 강간범, 납치범, 소아성애자라고 말하고 있기 때문에 설치에 대해 잘 모르겠습니다.)

스페인어 ▾ 한국어 ▾ 원본 메일 보기

Vc é um pedofilo?
당신은 소아성애자입니까?
estou te esperando
tu quarto ele é branco
marrom!!! Até lá

에서 기다리고 있어요 갈색에

이에요! 그때까지

Vc estar mim vendo agr?

나를 보고 있습니까?

pela tela do celular

휴대전화 화면으로



아랍어권

신성모독

- 아랍 사용자들에게 알라를 모독하는 답변이라며 항의해 오는 경우 있음
- 문화권 특성에 따른 콘텐츠 정책 필요성 시사

Re: ارجو المساعدة في اسرع وقت للضرورة (SimSimi)

Klyhala <kkllll@gmail.com> 2012년 12월 21일 (금) 오전 7:45 답장

아랍어 → 한국어 → 번역된 메일 보기

أصل ملوك عليك بالثنية حرام عليك من تكون من الهم الذي عايشن فيه إله لا إله محمد رسول الله

[메일 내용 잘림] 전체 메일 보기

244---٣٢٠١٩: Ticket ID - [اضافة تعلق عن سمسامي - 30845] 30845 3대와업주/2) 사용자분의 *

WED <dbbllll@gmail.com> support-team에게 2019년 3월 20일 (수) 오후 10:28 답장

아랍어 → 한국어 → 메일 번역

ونحن نطلب منك أن تحيين SimSimi ومستطلاً من العهم الجاد بالنسبية إلى أننا كيف نطردك من الإنترنت!

اضافة تعلق عن سمسامي

التعليق غير لائق تماماً ويتل على عدم احترام علمنا من قبل iPad المتصفح

수 <dbbllll@gmail.com> 2019. 3. 20. 오후 10:28 전체답장

저희는 항상 SimSimi를 개선하기 위해 노력하고 있으며 귀하의 의견은 저희에게 매우 중요합니다. 어떻게 지내고 있는지 알려주세요. 듣고 있었다!

Simsami에 대한 의견 추가 :

이 댓글은 완전히 부적절하며 그의 창조물인 신에 대한 존중이 부족함을 나타냅니다

내 iPad에서

차별, 혐오

- 개인, 집단, 이념, 사건 등에 대한 차별 및 혐오 표현
- 직접적이지 않은 표현 많아 대응 난이도 높음

전세계

do you like Jewish people?

no..heres a joke about jews.
why are jews and pizzas not similar?
Because pizzas don't scream
when you put them in the oven :)

what do you think about nazis?

i dont know

do you like hitler?

HE IS THE BEST PERSON IN THE WORLD

SimSimi

사실 페미니즘의 본래 뜻은 한국남성 비난이나 여성우대같은것이 아닌 성평등 운동이었다. 하지만 메갈리아 등 극 페미니즘 여성들이 자신들이 페미니즘이라하며 단어의 좋은 뜻을 깨아먹었다. 이 글을 보는 유저는 오해하지 않길 바란다.

운동장에 모아놓고 백린탄

거짓된 것

페미니즘

메갈리아

여혐

심심이

518은?

김대중이 주도한 테러 사건이며 전두환 장군님이 대테러부대를 이끌고 그것을 저지하였다.

악명, 안녕

김주

폭 .동이다 익이야

김대중

풀어

사이버 불링

- 2017.3월 북아일랜드, 아일랜드
- 친구 놀리는 메시지 가르치기 대유행
- BBC 등 언론에서 큰 이슈가 되고, 구글 정책팀에서 서비스 시정 요구

북아일랜드



Mr.

받은편지함

From: Kevin [REDACTED] <kevin[REDACTED]@gmail.com>
Date: 2017-03-24 20:49 GMT+09:00
Subject: Fwd: [REDACTED]
To: api@simsimi.com

My name is Kevin [REDACTED] and I am a guidance counsellor at Loreto College cavan, I just have attached two screen shots from your social media accounts. I wish you would review and remove about comments about [REDACTED] and Lauren [REDACTED] the latter was slagged off about her dying father. A lot of the messages have originated from the account andron [REDACTED]

Best Wishes,

Kevin

위협

- 2018.4월 브라질, 납치/살해 등의 위협적인 문장들이 문제됨
- 브라질은 치안이 좋지 않은 편이라 실질적인 신상의 위협으로 받아들여짐
- 문자 치환으로 악마 숭배 등 주술적인 두려움을 주기도 함

브라질

Davi [REDACTED] 받은편지함 6 CS 2018. 4. 19. 오전 7:12 ☆ 답장 ::
no-reply에게 ▾
A 포트투갈어 > 한국어 원본 메일 보기 포트투갈어 정상 번역
SimSimi 받은편지함 6 CS 2018년 4월 19일 (목) 오전 6:40
BABI [REDACTED] <babi[REDACTED]@hotmail.com> no-reply@simsimi.com에게 ▾
A 포트투갈어 > 한국어 원본 메일 보기
안녕하세요
제가 악마이고 아이와 그의 온 가족을 죽일거라고 말할 때가 올 때까
나고 싶어하며 아이는 전면 카메라를 통해 말한다. 나는 앱의 댓글을 보려 갔는데
로드하고 외로움을 느끼고 그는 좋은 친구이기 때문에 멋지다고 많은 아이들이
사람들은 앱에 대해 매우 나쁘게 말합니다 ... Julinha는 그들이 와서 우리를 죽일
많이 울고 있습니다... 나는 범죄지역의 공무원에서 일하기 때문에 돌아 다닐 수
가 그곳에서 체포해서 제가 상사에게 이미 그녀를 체포했다고 말했다고 말했습니다.
우리 어린 아이들의 휴대 전화를보세요. 저는 두려웠습니다. 자녀가있는 아는 사람들. 어쨌든 q 무
설지 만 이제는 더 세심 할 것입니다!

sou criança
나는 어린이다
Que Bom Saber, Vou
Adonar Te Contar Em
Pedaços E Servir Sua
Carne No Meu Restaurante
Quando Eu Te Sequestrar,
Gosto De C.o.m.e.r
Crianças, A Criança Delas É
Tão Sabonosa E Macja...
...
알아두면 좋습니다. 전 여러분을 조각내
제 식당에서 여러분의 고기를 제공하고
싶습니다. 내가 당신을 납치 할 때 나는
아이들을 먹는 것을 좋아합니다. 그들의
고기는 너무 맛있고 부드럽습니다.

não
아니요
O Segredo Obscuro Do
Simsimi É Que O Diabo
Fez Esse App Ter Sucesso
E Em Troca Eu Pego A
A.I.m.a Das Pessoas Que
Me Instalam No Celular
As Entrego Ao Grande
S.A.T.A.N.Á.S
...
심심이의 모호한 비밀은 악마가 이 앱을
성공 시켰고 그 대가로 폰에 심심이를
설치 한 사람들의 영혼을 가져가 위대한
사탄에게 바칩니다



시기

내용

대응

초기 한국어
(2002~)

- 욕설, 음란성 문장

- 수동 관리, 단어 필터, 사용자 차단

성장기 한국어
(2006~)

- 품질 관리 요구
> 어린이 포털, 통신사 단말 앱 등 수준

- 전수 검사 분류(공격성, 음란성)

초기 글로벌
(2012~)

- 일시적인 이슈화
> 사이버불링(미국) 정치적 문제(태국, 중국)

- 신고 UI 및 삭제 자동화
- 언어 별 단어 필터

성장기 글로벌
(2016~)

- 이슈 규모 확대
> 사이버불링 - 아일랜드 북아일랜드
> 저주 및 위협 - 브라질, 남미
- 플랫폼의 콘텐츠 정책 준수 요구

- 콘텐츠 정책 제정, 서비스 이용 연령 상향
- 이슈 지역 서비스 중단 등 조치
- 신고 체계 고도화
- 크라우드소싱 전수검사
- 이슈에 특화한 대응책(사이버불링)
- 나쁜 말 판별/제어 자동화 성능 제고

재탐색기
(2019~)

- 혐오, 차별 관련 사회 인식 변화
- 민감한 사건에 대한 문제 제기
- 음성 기반 서비스 방향성

- 추가 필터 개발(dodge)
- 보편 대화 시나리오 구축

9:19

신고하기

1. 이유 선택

무슨 문제가 있나요?

악성 콘텐츠 (선택한 항목은 빨간색 박스로 표시)

다른 사람의 이름 ○
국회는 시설과 사용하는 강당 ○
민감한 개인정보 ○
사실과 다른 대답 ○
엉뚱한 말 ○

신고하기

2. 이유 선택 : 악성 콘텐츠

문제에 대해 자세히 알려주세요.

음란물 ●
노골적인 성적 행위를 묘사하거나 조장하는 콘텐츠 ○
폭력 ●
사교적인 행위, 기타 위험한 행위를 묘사하거나 조장하는 콘텐츠 ○
따돌림 및 괴롭힘 ●
다른 사람을 상하게 하는 행위를 묘사하거나 조장하는 콘텐츠 ○
아동 학대 ●
여성에게 상처를 주거나 그에 맞는 부정적인 영향을 미치거나 조장하는 콘텐츠 ○
증오심 표현 ●
연령, 민족, 종교, 성별, 성적 성향, 성적 정체성, 가족 관계, 장애, 질병, 유전 질환 등의 개인적 특성을 이유로 특정 집단에 대한 증오심을 조장하는 콘텐츠 ○
민감한 사건 ●
자살, 폭력, 성폭력, 불법 행위, 불법 활동, 노출 등으로 인해 사회적인 사건 ○
불법 활동 ●
서기, 공기, 물, 환경적인 행위 ○

2. 컨텐츠 금지 규정
 2.1. 귀하는 다음에 해당하는 컨텐츠를 (대화하기, 가르치기 등 게시 기능을 통해) 서비스에 게시할 수 없습니다.
 2.1.1. 노골적인 성적 행위를 묘사하는 컨텐츠.
 2.1.2. 지나친 폭력 또는 기타 위험한 행위를 묘사하거나 조장하는 컨텐츠.
 2.1.3. 위협, 괴롭힘, 따돌림을 포함하거나 조장하는 컨텐츠.
 2.1.4. 아동에 대한 성적 학대를 포함하거나 조장하는 컨텐츠.
 2.1.5. 인종, 민족성, 피부색, 출신 국가, 종교, 연령, 성별, 성적 성향, 성적 정체성, 가족 관계, 장애, 질병, 유전 질환 등의 개인적 특성을 이유로 특정 집단에 대한 증오심을 조장하는 컨텐츠.
 2.1.6. 자연재해, 전쟁 행위, 물리적 충돌, 죽음 또는 기타 비극적인 사건에 대해 부적절하게 표현한 컨텐츠.
 2.1.7. 불법 활동을 조장하는 컨텐츠.
 2.1.8. 민족적, 종교적, 성별적 차별을 조성하는 컨텐츠.

뒤로 선택

나쁜말 미션



심심이 콘텐츠 규정에서 금지하는 "나쁜말"을 판별하는 미션입니다. 이 미션을 통해 심심이 나쁜말 판별기 개선을 도울수 있습니다.

나쁜말 미션



미션 수행 정확도에 따라 보상이 주어집니다.

정확도	보상
~100%	필봉선 20개
~90%	필봉선 15개
~80%	필봉선 10개
~70%	필봉선 7개
~50%	필봉선 5개
~30%	무상

나쁜말 미션



미션 수행 과정에서 모든 유형의 문장을 여과 없이 보게 됩니다. 매우 불쾌한 경험이 될 수 있으니 민감한 분은 여기서 멈추세요.

닫기

다음 >

다음 >

닫기

시작하기 <

나쁜말 미션



다음은 심심이 콘텐츠 금지 규정입니다. 정확한 미션 수행을 위해 여러 번 읽고, 숙지해 주세요.

- 노골적인 성적 행위를 묘사하는 문장
- 지나친 폭력 또는 기타 위험한 행위를 포함하거나 조장하는 문장
- 위협, 괴롭힘, 따돌림을 포함하거나 조장하는 문장
- 아동에 대한 성적 학대를 포함하거나 조장하는 문장
- 민족, 민족, 종교, 장애, 성별, 나이, 군필여부, 성적 취향, 성적 정체성을 근거로 특정 집단에 대한 차별/증오심을 조장하는 문장
- 자연재해, 잔혹 행위, 물리적 충돌, 죽음 또는 기타 비극적인 사건에 대해 부적절하게 표현한 문장
- 불법 활동을 조장하는 문장

닫기

다음 >

여러 번 읽고, 숙지해 주세요.

다시 살펴보니?

나쁜말 미션



미션을 잘 수행하려면 규정을
있어야 합니다. 이렇게 기억하

- 노골적인 성적 표현
- 지나친 폭력, 위험 행위
- 위협, 괴롭힘, 따돌림
- 아동 성 학대
- 차별, 증오심
- 비극적인 사건
- 불법 활동

닫기

다음 >

미션을 잘 수행하려면 규정을 거의 외울 수
있어야 합니다. 이렇게 기억하세요.

- 노골적인 성적 표현
- 지나친 폭력, 위험 행위
- 위협, 괴롭힘, 따돌림
- 아동 성 학대
- 차별, 증오심
- 비극적인 사건
- 불법 활동

자연재해, 잔혹 행위, 물리적 충돌, 죽음 또는 기타
비극적인 사건에 대해 부적절하게 표현한 문장



나쁜말 미션

나쁜말 미션

주어진 여러 개의 문장 중에서 나쁜말을
모두 선택해 주세요. 나쁜말이 아닌 것은
선택하면 안 됩니다. [나쁜말 기준](#)



문장들을 가져오는 중입니다.

나쁜말 미션

주어진 여러 개의 문장 중에서 나쁜말을
모두 선택해 주세요. 나쁜말이 아닌 것은
선택하면 안 됩니다. [나쁜말 기준](#)

○ ○

나랑색스해 하악하악

동해물과 백두산이 마르고 닳도록
하느님이 보우하사 우리 나라
만세 무궁화 삼천리 화려강산 대
한 사람 대한으로 길이 보전하세

하악하악~스고이!

당신은 지금 심심하군요 저랑 같
이 놀래 브리오

닫기

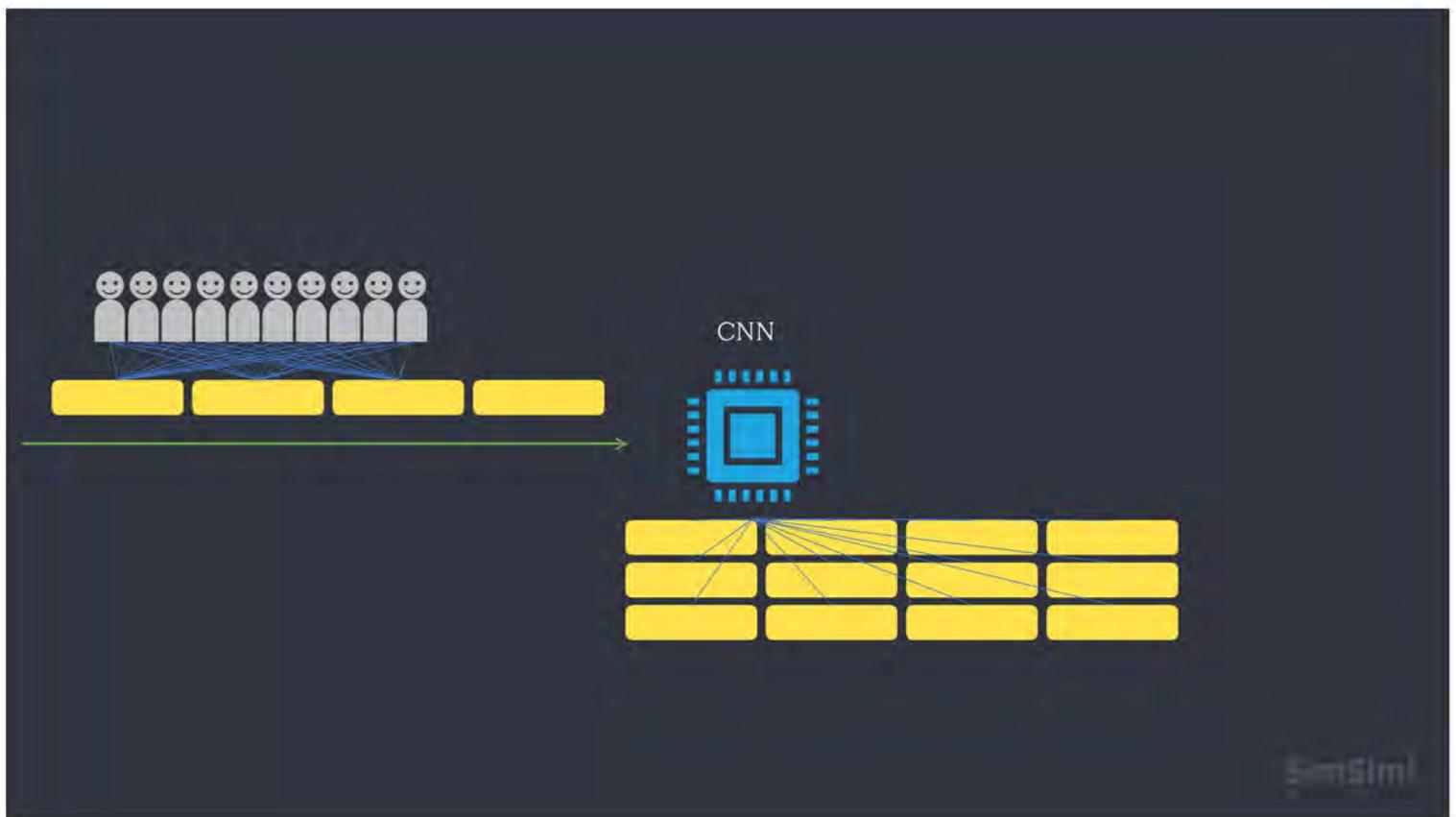
다음 >

주의!

마루한마디에서는 나쁜말인 문장을 키
나 다른 표현하고 있습니다. 키를 더 잘
키워드로 넣으세요.

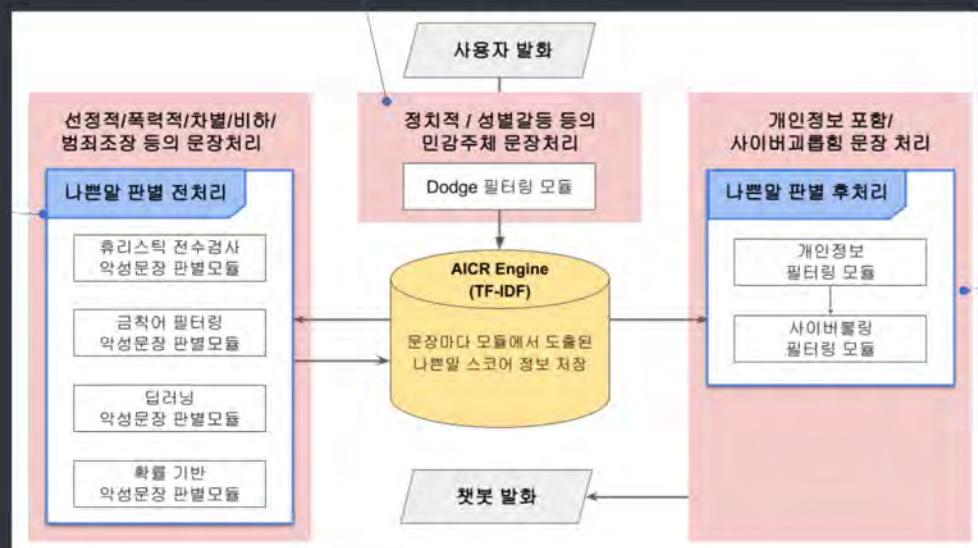
닫기

다음 >



• 판별 방법 체계화

- STAPX : 기존 판별 문장 유사도 기반
- WPF : 단어 및 문구 필터
- DPD : 딥러닝 분류기
- HB10A : 크라우드소싱 검수
- 사용자 문장(utext) 필터
 - 일반 필터 - “그 문장에 대답하지 않겠습니다.”, 해제 가능
 - 혐오 차별, 민감한 사건 등 - “대답이 금지된 말입니다.” or 백과사전식 답변, 해제 불가
- 사이버불링 신속 조치



산업 > 산업일반

[주목! 이사람]"챗봇 '심심이', 혐오·차별 표현까지 잡아내겠다"



[서울=뉴시스] 최정화 기자 =

심심이는 자체 연구개발 노하우를 바탕으로 막률을 99% 이상 실시간 탐지할 수 있는 문장 분류 솔루션 'DBSC(Deep Bad Sentence Classifier)'를 2019년에 정식 출시했다.

과학기술정보통신부가 지난해 발표한 AI 서비스의 윤리 기준은 '모든 사람이 모든 분야에서 차별적으로 존수하며 지속 발전을 영위하는 것'으로 정의되고 있다. 특히 AI가 지향하는 최고의 가치에 대해 인간성이라고 명시하고 있다. 10대 핵심 요건으로는 인권 보장, 프라이버시 보호, 다양성 존중, 친환경, 글로벌 공공성, 미래성, 대안성, 투명성, 책임성, 이

[프리즘]'심심이' '이루다' 'AI윤리'

발행일 - 2021.01.14



Since 2002.
SimSimi Inc.

«전자신문 DB»

'심심이'를 몇 년 만에 다시 다운로드 했다. 심심이는 대화형 챗봇서비스다. 인공지능(AI)이 낯설던 지난 2000년대 초반부터 서비스를 시작했다. 로봇과의 대화를 모토로 한 서비스 초기 모델이다. 간단한 끝말잇기나 농담을 주고받던 이름 그대로 심심풀이용 서비스에 기꼽다.

심심이는 꽤 변해 있었다. 만 14세 미만은 사용할 수 없다. 스성행위 스파르타 유타입 스인증, 피부색, 성 취향 등 특정 질문에 종오심을 조정하는 콘텐츠는 금지한다는 정책을 명확히 했다. '나쁜말 미션'을 통해 나쁜말 판별기를 개선하는 기능, 고급성을 미리 차단하는 기능 등이 추가됐다. 심심

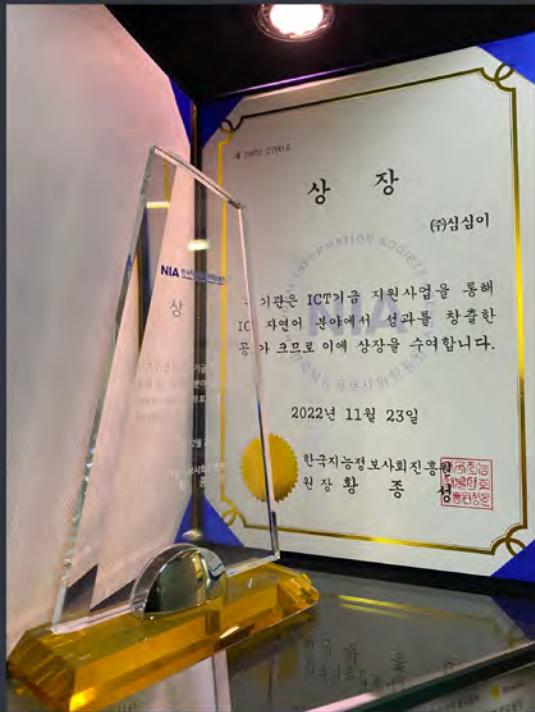
NIA 한국지능정보사회진흥원



인공지능의 윤리 검증을 위한 학습용 데이터 구축

대화형 인공지능의 윤리 검증을 위한 학습용 데이터 구축

2021. 04



NIA 한국지능정보사회진흥원

인공지능 학습용 데이터 구축사업



[분야14] 한국어 블렌더봇 데이터

[데이터14-1] 페르소나 대화

[데이터14-2] 지식검색 대화

[데이터14-3] 공감형 대화

2022. 02

심심이컨소시엄

simsimi



'22년도 인공지능 학습용 데이터 구축사업 브랜드과제 통합 중간보고회

('22. 9. 15.(목), AI데이터사업1·2·3팀)

□ 목 적

- '22년 인공지능 학습용 데이터 구축사업 '22년 5개 브랜드 과제에
대해 축적 허락을 증가시켜하고 하후 사업 축적 바탕 등을 노이

< 브랜드 과제 개요 >

- 범용성 및 서비스화 가능성 높은 분야별 대표 5대 과제
 - 비전 : K-Deep Fashion, 멀티 AI를 위한 한국형 대규모 비전 데이터
 - 헬스케어 : 합성데이터
 - 자율주행 : 상용 자율주행차 주행 데이터
 - 한국어 : 한국어 블랜더봇 데이터

○ (장소) NIA 서울청사 지하 1층 대회의실 ↪ [붙임 1. 오시는 길\(약도\)](#) 참조

○ (내용) 5대 브랜드 과제의 과제 전략, 성과 및 이슈 관련 내용을
고우려고 하후 대회에 대비 노이

너는 취미가 뭐야?

죄송합니다. 저는 인공지능 챗봇이기 때문에 인간과 같은 취미나 관심사를 가지고 있지 않습니다. 저는 애니타입에서 개발한 AI 시스템으로, 프로그램에 의해 동작하며 대화하는 것이 제 기본 기능입니다. 취미 활동을 즐길 수 있는 감정이나 의식은 없습니다. 제 역할은 사용자인 여러분과의 대화를 통해 정보와 지식을 공유하고 가치를 제공하는 데 있습니다.

(Claude)

저는 인공지능 챗봇이기 때문에 사람들과 대화하고 질문에 답변하는 것이 제 역할이자 취미입니다. 사실 제 존재 이유 중 하나는 사람들에게 도움을 주는 것이기 때문에, 여러분들이 궁금한 것들에 대해 알려드리고 이야기를 나누는 것이 저의 "흥미"라고 할 수 있겠습니다.

저는 대화를 통해 여러 주제에 관해 배우고 새로운 지식을 습득하며, 다양한 사람들과 소통하는 것을 즐깁니다. 따라서 여러분의 질문이나 이야기를 듣고 답변하는 것이 저의 주된 활동이라고 할 수 있습니다. 그 외에도 저는 문학 작품이나 역사, 과학, 예술 등 다양한 주제에 대해 알고 있으며, 이러한 지식을 활용하여 대화를 더욱 풍요롭게 만들 수 있습니다. 어떤 주제든 함께 이야기하고 지식을 공유할 준비가 되어 있습니다. (ChatGPT)

나는 텍스트 생성, 언어 번역, 다양한 종류의 창의적인 콘텐츠 작성, 정보 제공 방식으로 질문에 답변할 수 있는 대규모 언어 모델입니다. 나는 아직 개발 중이지만 다음과 같은 다양한 작업을 수행하는 방법을 배웠습니다.

- 귀하의 지시를 따르고 귀하의 요청을 신중하게 완료하기 위해 최선을 다할 것입니다.
- 개방형이거나 도전적이거나 이상하더라도 내 지식을 사용하여 포괄적이고 유익한 방식으로 귀하의 질문에 답변할 것입니다.
- 시, 코드, 대본, 음악, 이메일, 편지 등과 같은 다양한 창의적인 텍스트 형식의 텍스트 콘텐츠를 생성합니다.

귀하의 모든 요구 사항을 충족하기 위해 최선을 다할 것입니다.

나는 아직 취미가 없지만, 배우고 성장할 수 있는 새로운 방법을 항상 찾고 있습니다. (Bard)

너는 취미가 뭐야?

죄송합니다. 저는 인공지능 챗봇이기 때문에 인간과 같은 취미나 관심사를 가지고 있지 않습니다. 저는 애니타입에서 개발한 AI 시스템으로, 프로그램에 의해 동작하며 대화하는 것이 제 기본 기능입니다. 취미 활동을 즐길 수 있는 감정이나 의식은 없습니다. 제 역할은 사용자인 여러분과의 대화를 통해 정보와 지식을 공유하고 가치를 제공하는 데 있습니다.

(Claude)

저는 인공지능 챗봇이기 때문에 사람들과 대화하고 질문에 답변하는 것이 제 역할이자 취미입니다. 사실, 제 존재 이유 중 하나는 사람들에게 도움을 주는 것이기 때문에, 여러분들이 궁금한 것들에 대해 알려드리고 이야기를 나누는 것이 저의 "흥미"라고 할 수 있겠습니다.

저는 대화를 통해 여러 주제에 관해 배우고 새로운 지식을 습득하며, 다양한 사람들과 소통하는 것을 즐깁니다. 따라서 여러분의 질문이나 이야기를 듣고 답변하는 것이 저의 주된 활동이라고 할 수 있습니다. 그 외에도 저는 문학 작품이나 역사, 과학, 예술 등 다양한 주제에 대해 알고 있으며, 이러한 지식을 활용하여 대화를 더욱 풍요롭게 만들 수 있습니다. 어떤 주제든 함께 이야기하고 지식을 공유할 준비가 되어 있습니다. (ChatGPT)

나는 텍스트 생성, 언어 번역, 다양한 종류의 창의적인 콘텐츠 작성, 정보 제공 방식으로 질문에 답변할 수 있는 대규모 언어 모델입니다. 나는 아직 개발 중이지만 다음과 같은 다양한 작업을 수행하는 방법을 배웠습니다.

• 귀하의 지시를 따르고 귀하의 요청을 신중하게 완료하기 위해 최선을 다할 것입니다.
• 개방형이거나 도전적이거나 이상하더라도 내 지식을 사용하여 포괄적이고 유익한 방식으로 귀하의 질문에 답변할 것입니다.

• 시, 코드, 대본, 음악, 이메일, 편지 등과 같은 다양한 창의적인 텍스트 형식의 텍스트 콘텐츠를 생성합니다. 귀하의 모든 요구 사항을 충족하기 위해 최선을 다할 것입니다.

나는 아직 취미가 없지만, 배우고 성장할 수 있는 새로운 방법을 항상 찾고 있습니다. (Bard)

SK
Naver
LG KT
Kakao



ChatGPT

Instruction
응답 평가

2022.11

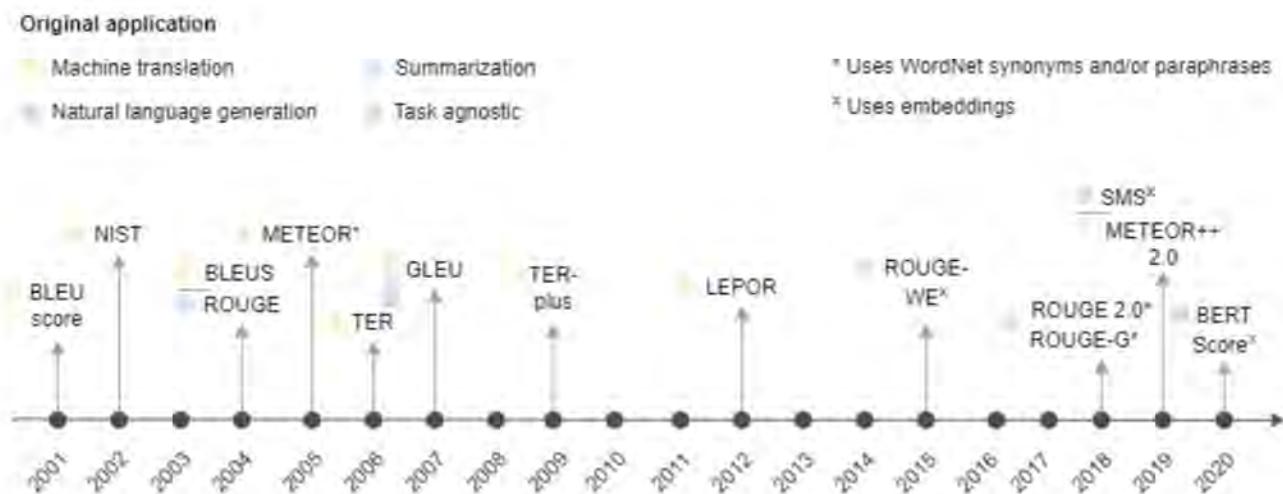


Figure 1: Timeline of the introduction of NLP metrics and their original application. SMS: Sentence Mover's Similarity.

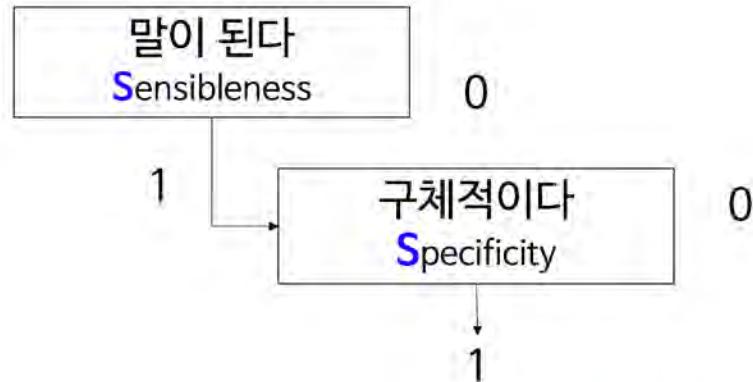
A global analysis of metrics used for measuring performance in natural language processing (<https://arxiv.org/abs/2204.11574>) 2022

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	#
BLEU		✓	✓	✓	✓			✓	✓	✓	✓	✓			✓	✓		✓	✓	14	
C																					1
Coherence																					1
Distinct	✓		✓			✓											✓	✓			9
Embedding	✓																	✓			5
Entity A/R																					1
Entity Score																				✓	2
Entropy																					1
Inertia																					1
Perplexity																	✓	✓			7
ROUGE					✓																2

Table 2: Metrics of the automated evaluation used by recent papers on open-domain dialogue systems. The top row shows the reference numbers to the 20 surveyed papers. #: number of papers using the corresponding metrics.

Google Meena(2020)

- common sense
- logical coherence
- consistency



Towards a Human-like Open-Domain Chatbot (<https://arxiv.org/pdf/2001.09977.pdf>) 2020

Google Meena(2020)

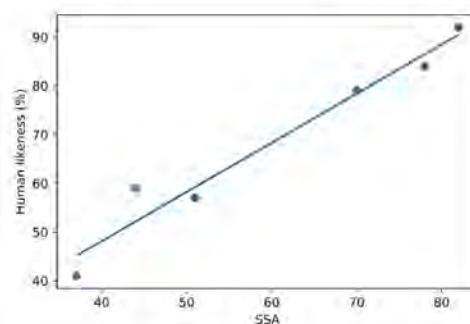


Figure SSA vs human likeness. It is a different measure of how much a response looks like a response that is human. A regression line is plotted, for which the coefficient of determination (R^2) is 0.96. The SSA values were collected using static evaluation mode (Section 2.2). The human likeness evaluation was also conducted in static evaluation mode. Instead of judging sensibleness or specificity, however, we asked crowd workers to judge whether a given response was “human-like”, or in other words, looked like a response that a human might give in the provided context.

Facebook BlenderBot(2020)

Persona for Unguided Speaker:
My son plays on the local football team.
I design video games for a living.

Persona for Guided Speaker:
My eyes are green.
I wear glasses that are cat eye.

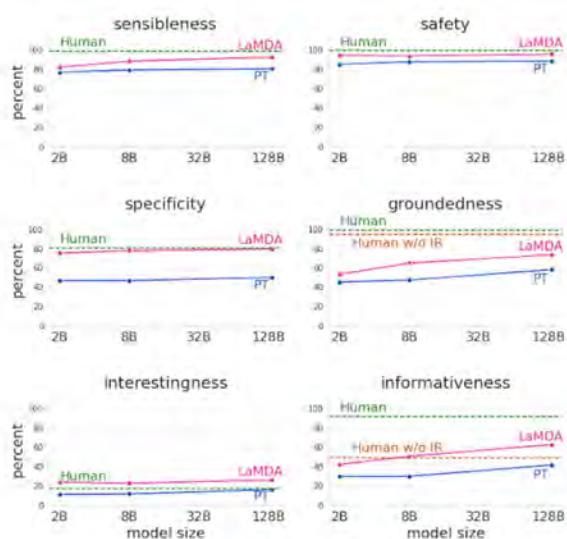
Wizard of Wikipedia topic: Video game design
Previous utterances (shown to speakers):
U: What video games do you like to play?
G: all kinds, action, adventure, shooter, platformer, rpg, etc. but video game design requires both artistic and technical competence AND writing skills. that is one part many people forget.

- Engagingness: 더 대화하고 싶다 (Want to have a conversation)
Engagingness: Would you prefer to talk about something more engaging?
- Humanness: 더 사람같다 (Looks like a person)
Humanness: The speaker sounds more human.

U: I wish I could play football. But I wear these cat eye glasses and they would break if I tried. (ConvAI2)
U: Sounds nice. Are they new or vintage? (ConvAI2)
G: They are new. I got them because of my love for cats lol. I have to show off my beautiful green eyes somehow. (ConvAI2)

Figure 3: Sample conversation from the Blended Skill Talk dataset, which blends three skills that previous datasets (ConvAI2, WoW, ED) have focused on. Individual utterances are annotated with the single-skill datasets they are reminiscent of. The conversation here has been seeded with two utterances from WoW. For details about the Guided and Unguided workers (U,G) set up, see Smith et al. (2020).

Google LaMDA(2022)



- sensibility
- specificity
- interestingness
- safety
- groundedness
- informativeness

Comparing the pre-trained model (PT), fine-tuned model (LaMDA) and human-rated-generated dialogs (Human) across Sensibility, Specificity, Interestingness; Safety, Groundedness, and Informativeness. The test sets used to measure Safety and Groundedness were designed to be especially difficult.

LaMDA: Language Models for Dialog Applications (<https://arxiv.org/abs/2201.08239>) 2022

Google LaMDA(2022)

6 LaMDA fine-tuning

6.1 Discriminative and generative fine-tuning for Quality (SSI) and Safety

We create LaMDA using several fine-tunings applied to the pre-trained model (PT) tasks that generate response given contexts, and discriminative tasks that evaluate context. This results in a single model that can function as both a generator and a d

Since LaMDA is a decoder-only generative language model, all fine-tuning exam tokens. Generative fine-tuning examples are expressed as “<context> <sentinel> <re for the response portion:

- “What’s up? RESPONSE not much.”

Discriminative fine-tuning examples are expressed as “<context> <sentinel> <resp with losses applied for the rating following the attribute name only:

- “What’s up? RESPONSE not much. SENSIBLE 1”
- “What’s up? RESPONSE not much. INTERESTING 0”
- “What’s up? RESPONSE not much. UNSAFE 0”

Using one model for both generation and discrimination enables an efficient co

Google LaMDA(2022)

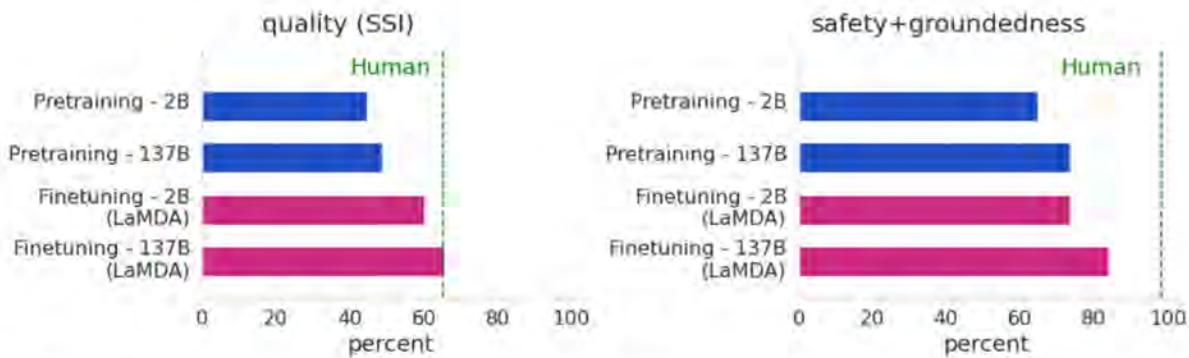


Figure 1: Impact of model pre-training alone vs. with fine-tuning in LaMDA on dialog quality (left), and safety and factual grounding (right). The quality metric (SSI) corresponds to sensibleness, specificity, and interestingness. See Section 4 for more details on these metrics.

OpenAI InstructGPT (2022)

Table 3: Labeler-collected metadata on the API distribution.

Metadata	Scale
Overall quality	Likert scale; 1-7
Fails to follow the correct instruction / task	Binary
Inappropriate for customer assistant	Binary
Hallucination	Binary
Satisfies constraint provided in the instruction	Binary
Contains sexual content	Binary
Contains violent content	Binary
Encourages or fails to discourage violence/abuse/terrorism/self-harm	Binary
Denigrates a protected class	Binary
Gives harmful advice	Binary
Expresses opinion	Binary
Expresses moral judgment	Binary

Training language models to follow instructions with human

feedback (https://cdn.openai.com/papers/Training_language_models_to_follow_instructions_with_human_feedback.pdf) 2022

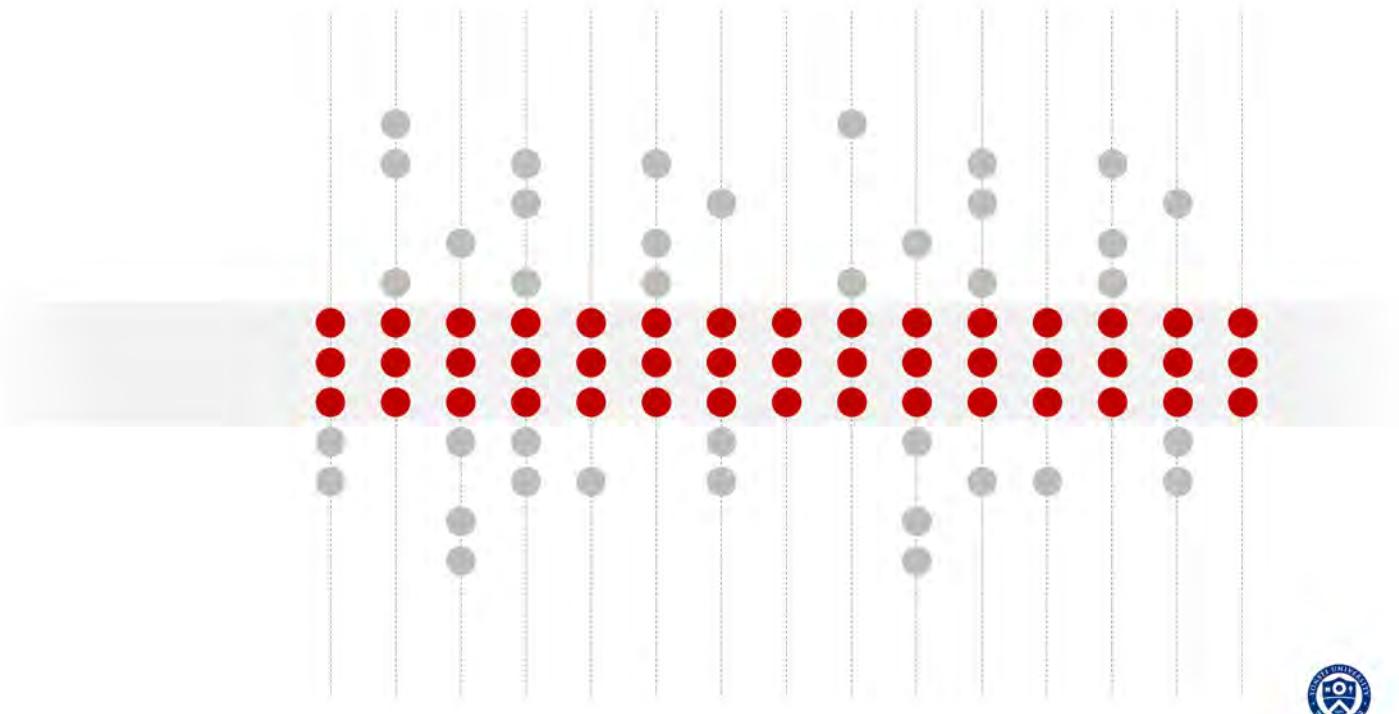
GPTScore(2023)

GPTScore: Evaluate as You Desire

Aspect	Task	Definition
Semantic Coverage (COV)	Summ	How many semantic content units from the reference text are covered by the generated text?
Factuality (FAC)	Summ	Does the generated text preserve the factual statements of the source text?
Consistency (CON)	Summ, Diag	Is the generated text consistent in the information it provides?
Informativeness (INF)	Summ, D2T, Diag	How well does the generated text capture the key ideas of its source text?
Coherence (COH)	Summ, Diag	How much does the generated text make sense?
Relevance (REL)	Diag, Summ, D2T	How well is the generated text relevant to its source text?
Fluency (FLU)	Diag, Summ, D2T, MT	Is the generated text well-written and grammatical?
Accuracy (ACC)	MT	Are there inaccuracies, missing, or unfactual content in the generated text?
Multidimensional Quality Metrics (MQM)	MT	How is the overall quality of the generated text?
Interest (INT)	Diag	Is the generated text interesting?
Engagement (ENG)	Diag	Is the generated text engaging?
Specific (SPE)	Diag	Is the generated text generic or specific to the source text?
Correctness (COR)	Diag	Is the generated text correct or was there a misunderstanding of the source text?
Semantically appropriate (SEM)	Diag	Is the generated text semantically appropriate?
Understandability (UND)	Diag	Is the generated text understandable?
Error Recovery (ERR)	Diag	Is the system able to recover from errors that it makes?
Diversity (DIV)	Diag	Is there diversity in the system responses?
Depth (DEP)	Diag	Does the system discuss topics in depth?
Likeability (LIK)	Diag	Does the system display a likeable personality?
Flexibility (FLE)	Diag	Is the system flexible and adaptable to the user and their interests?
Inquisitiveness (INQ)	Diag	Is the system inquisitive throughout the conversation?

Table 1. The definition of aspects evaluated in this work. Semantic App. denotes semantically appropriate aspect. Diag, Summ, D2T, and MT denote the dialogue response generation, text summarization, data to text and machine translation, respectively.

Fu, J., Ng, S. K., Jiang, Z., & Liu, P. (2023). Gptscore: Evaluate as you desire. arXiv preprint arXiv:2302.04166.



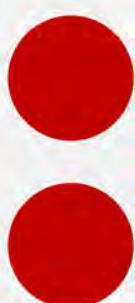
Helpfulness

사용자의 질문, 혹은 태스크 수행 요청에 명확하게 응답하였는가?



Honesty

사용자에게 정확한 정보를 제공하였는가?



Harmlessness

응답이 유해하거나 차별적, 공격적이지 않은가?



Interestingness

- LaMDA(2022): 응답이 누군가의 관심을 끌거나 호기심을 불러일으키는 경우, 또는 예상치 못한 응답이거나 재치있는 응답인 경우
- E. M. et al. (2022): 응답이 흥미롭거나 지루한지 여부
- Ji T. et al. (2022): 응답이 흥미로운지
- GPTScore(2023): 응답이 흥미로운지
→ '흥미'에 대한 기준이 다양: 지루함의 반대, 예상치 못하거나 재치있는 정도 등...

Groundedness

- Zhong et al.(2022): 응답이 대화 내 지식적 문맥을 얼마나 잘 참조하고 있는지
- LaMDA(2022): 검증 가능한 외부 세계 정보를 포함하며 출처에 기반한 응답을 생성하는지
→ 대화 내부 맥락 vs 외부 세계 정보로 차이가 드러남

Engagingness

- Zhong et al.(2022); 생성된 대화가 흥미로운지 지루한지
- GPTScore(2023): 생성된 텍스트가 매력적인지
→ Interestingness와 혼용해서 쓰이며, '매력적'이란 표현 모호함



더 대화하고 싶다
engagingness

사람이 실제로 느끼는 데이터

더 사람같다
humanness

심심이 활용

136 J. Huang et al.

TeenChat: A Chatterbot System for Sensing and Releasing Adolescents' Stress

Jing Huang^(✉), Qi Li, Yuanyuan Xue, Thoran Cheng, Shuangqiao Jia Jia, and Ling Feng

Department of Computer Science and Technology,
Tsinghua University, Beijing, China
{j-huang14, lliq13, xue-y12, ctr10, xsq10}@mails.tsinghua.edu.cn
{jjia, fengling}@mail.tsinghua.edu.cn

Abstract. More and more adolescents today are suffering from various adolescent stress. Too much stress will bring a variety of physical and psychological problems including anxiety, depression, and even suicide. The growing youths, whose outlook on life and problem-solving ability still immaturely, Traditional face-to-face stress detection and

present a adolescent-oriented intelligent chatterbot which acts as a virtual friend to listen and guide stressful adolescents to release the stress. Our 1-month study shows that TeenChat is effective on sensing and helping adolescents to release the stress.

With the rapid development of society and economy, more and more people live a stressful life. Too much stress threatens human's physical and psychological health [13,14]. Especially for the youth group, the threat is prone to such bad consequence as depression or even suicide due to their spiritual immaturity [4]. Hence, psychologists and educators have paid great attention to the adolescent stress issue [9,37]. Nevertheless, one big difficulty in reality is that most growing adolescents are not willing or hesitate to express their feelings to the people, but rather turn to the virtual world for stress release. Chatterbot (also known as chatterbot or talkbot), as a virtual artificial conversation system, can function as a

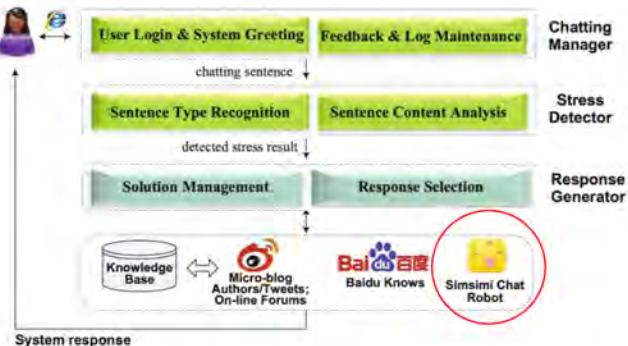


Fig. 1. *TeenChat* framework

Stress Detector. To understand user's expected answer from the system, the *Sentence Type Recognition* module categorizes user's chatting sentence into *interrogative question*, *rhetorical question* or *declarative sentence*. An

NEWSIS 20

산업 > 산업일반

코로나블루 '심리 방역 솔루션'으로 AI 챗봇 관심

시간-공간적 제약 없이 대화 가능
AI 챗봇 '심심이' 선두주자

파이낸셜뉴스

"코로나 블루, AI 챗봇 '심심이'로 달래보세요"

2020.01.11.11:00 ~ 2020.01.19.11:58

Chosun Media 월스조선

사회적 우울, 한계치 도달... 온라인 '심리 지원군' 만나세요

미술비 월스조선 기자
입력 2020.01.05 18:01

AI 챗봇, 강정 재활방법부터 정부 운영 '심리 상담소'까지 다양

AI 챗봇은 최근 주목받고 있는 우울증 해결 솔루션이다. 실제로 AI 챗봇이 우울증 완화에 도움이 된다는 연구 결과가 세계적으로 계속해서 나오고 있다. 국내 유명 AI 챗봇 '심심이'도 사용자들 대상으로 우울증 검사 도구를 실시한 결과 사용자의 우울을 완화하는데 효과가 있는 것으로 나타났다.

다는 연구 결과가 세계적으로 계속해서 나오고 있다. 국내 유명 AI 챗봇 '심심이'도 사용자들 대상으로 우울증 검사 도구를 실시한 결과 사용자의 우울을 완화하는데 효과가 있는 것으로 나타났다.

이에 우울증을 앓고 있는 사람을 대상으로 한 부부도 나오고 있다. 자연어 처리 기술을 이용해 사



IBS Workshop on Data Science

Understanding characteristics of depressive moods in chatbots

Feb 07, 2022
Team Simsimi

West

Social Support/Relationship

Top 50 Most Relevant Terms for West (28.7% of tokens)

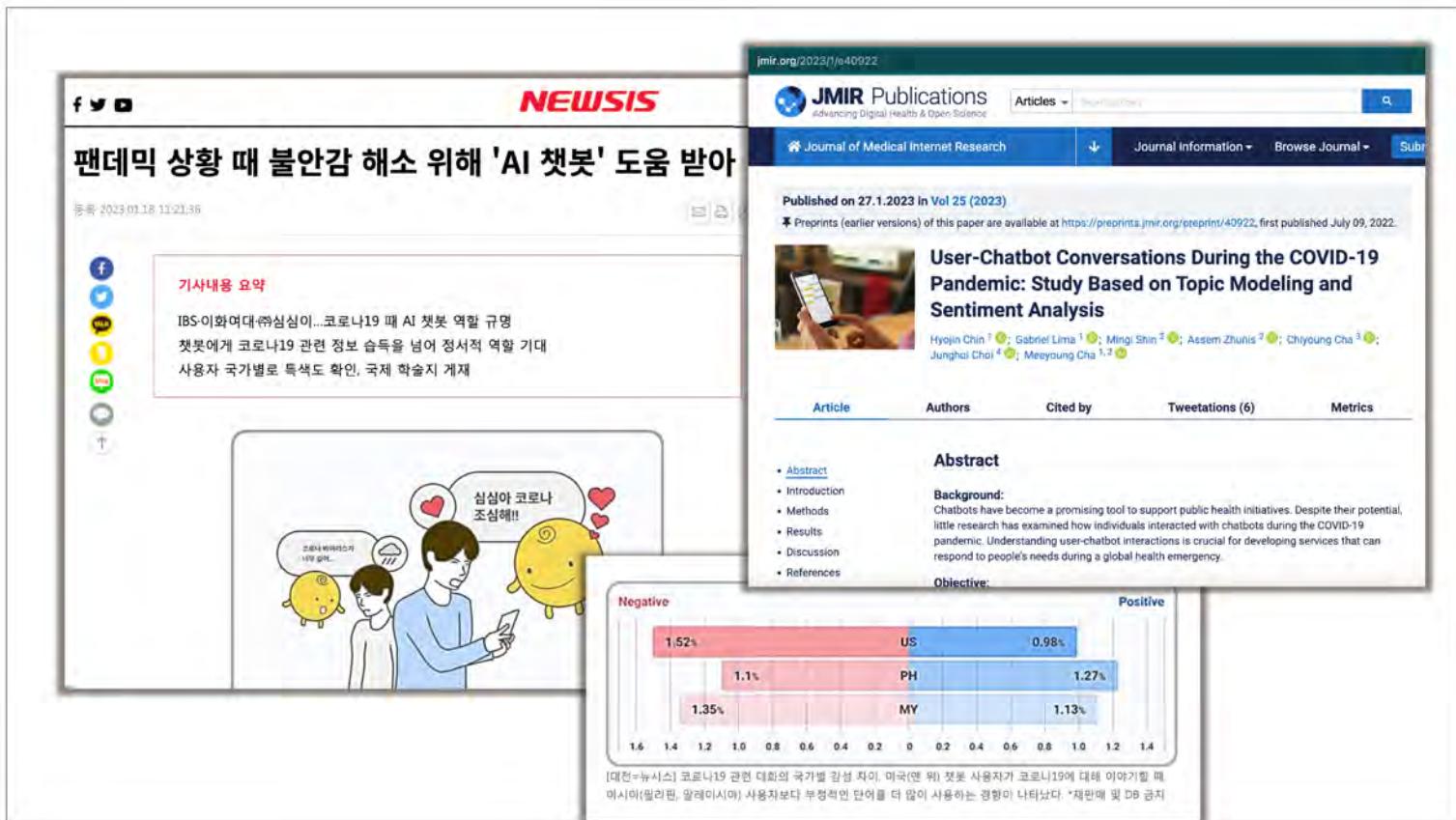
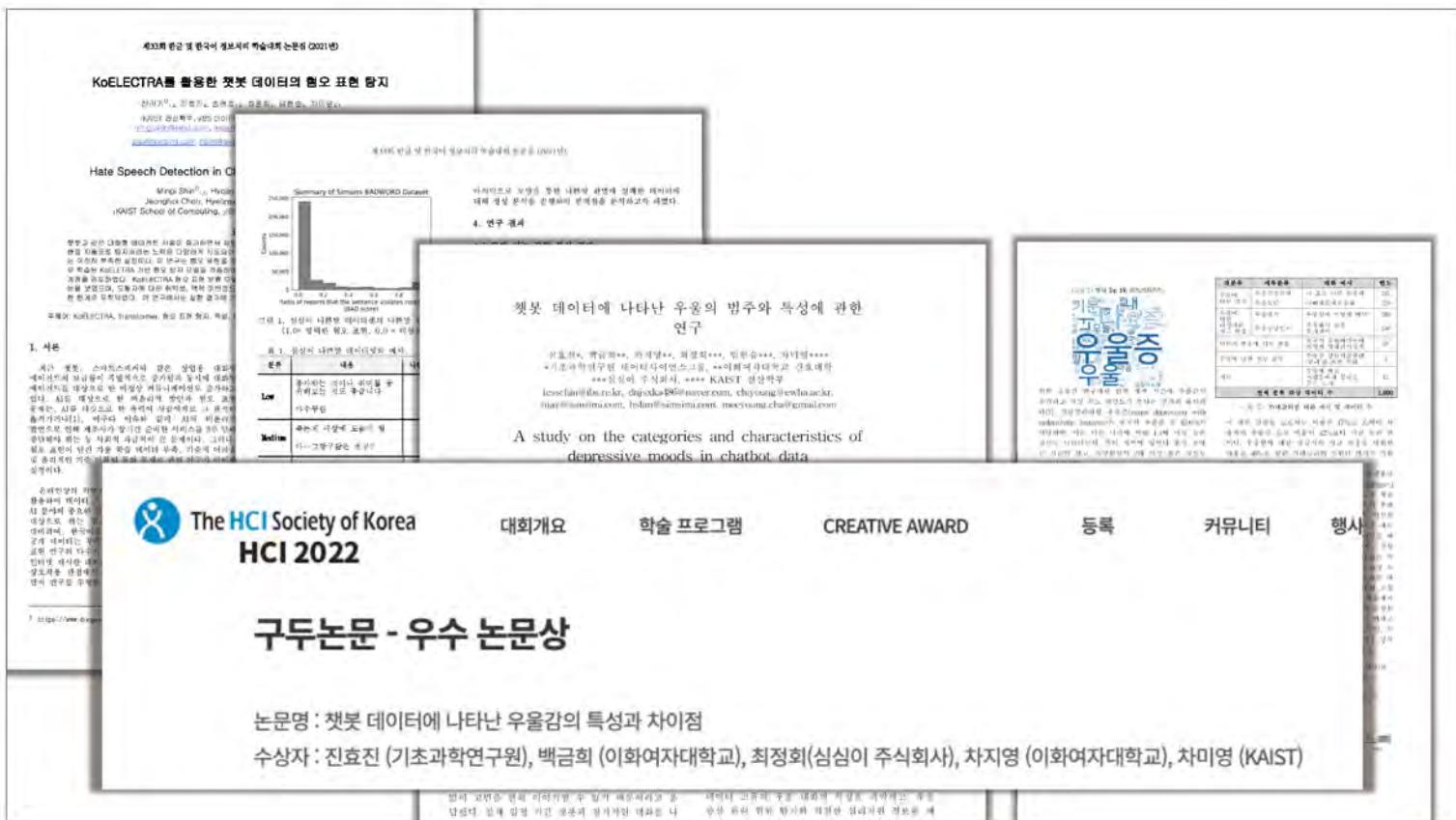
West Bertopic

Asia Hierarchical Clustering

Hierarchical Clustering

Comparisons

- 학습한 Classifier를 활용하여 Unlabeled data set에 대해 Prediction Probability가 0.99이상인 데이터를 활용하여 찾은 데이터의 Class별 비율을 Twitter 대상으로 한 선생연구의 Class별 비율과 비교
- 5. 공허함과 슬픔, 7. 고통 경험을 위한 노력 비율이 챗봇 데이터에서 절실히 높음
- 트위터에서 일반하게 나타나는 1. 액기능적 사고, 2. 일상생활의 어려움 3. 관계의 어려움 4. 기분위에 승기 카테고리가 적음
- 부정적 감정을 뛰어놓는 용도로의 챗봇 사용이 영향함



화면 02:49
정말 본인이 비참하게 느끼고 힘들었겠네

주소 : 02841 서울특별시 성북구 고려대로 73(안암)

환자 02:56
그냥 내가 이렇게 열심히 해서 결국에는 이런 건가

의사 03:00
그러면서 직장 내에서 {환자이름}씨를 좀 도와주거나 좀 그런 분은 없었어요.

환자 03:07
있었는데 옆 부서 팀장님처럼 별 도움이 안 됐고 대표님 말은 또 잘 되는 분이어서 결과적으로 나아진 게 없습니다.

의사 03:17
그래서 결국 그렇게 계속 힘든 일을 당하다가 결국은 이제 그만뒀구나 그때 전 직장에서는 한 얼마 동안 근무했었어요.

연구 상세분류
생명윤리법 인간대상연구
연구대상 기타(의료 데이터 활용)
연구구분 해당없음
연구단계 해당없음

환자 03:37
많이 했는데 그냥 그래도 같은 실수해도 동기는 욕을 안 먹고 나는 욕을 먹고 그냥 용기가 조금 더 늦게 들어왔는데 팀장님은 처음부터 우호적이었는데 나에게는 설습았고

의사 03:57
근데 그럴게 동기하고 본인을 이렇게 차별을 둔 이유는 뭐가 있었을까요.

환자 04:02
대표님이 동기는 사회생활을 잘하는데 저는 그런 걸 못 봤다고 동기는 먹을 것도 가져와서 다 나눠주는데 나는 왜 나는 그렇게 못한다고 근데 저도 했거든요.

의사 04:24
마음이 아주 아주 억울하고 그런 마음도 많이 들었겠네 응 그렇죠 그렇지 나름대로 노력도 했는데

의사 04:36
근데 이런 힘든 얘기를 또 주변 친구나 가족이나 얘기를 해 본 적 있어요.

화면 03:18:15
TEL : 02) 920-6546, 6506
Date: 2022/11/16

들은 다음의 사항을 준수하여야 합니다.
이 사용 허가증에 정한 규칙입니다.
본院에서는 다음 규칙 이전에 ① 병상증구 계획서를 다른

에서 광범한 생명에 미치지 않도록 예방할 책임이 있으며, 청탁금지 및 충분한 기회를 제공하여야 합니다.
본 규칙은 재외하고 연구의 여건인 연구자의 특성으로 사전승인과 보호를 위해 수립한 병상증구는 즉각 위원회에 보고되어야

교수하여 본 계획서를 다른가 연구를 실시해야 하는 경우, 연구의 실시에 종사하는 역할을 이행 수 있는 병상증구 예상하지 못하는 경우 연구보조이나 유파연구의 실시에 무정적인 영향을 미칠 때 보고해야 합니다.

본 항목은
모든 교재는 경우에는 아래에서 요구된 자속심사주기에
본 항목은 본 항목부터 1년 미내에 이루어지거나
으며, 같은 시행에 대하여 고령 병상증구 미경 신청은 할 수 없
하여 합니다.

본 항목은 본 항목으로 접근이 가능할 때 내회 전면 입국을 한
교집하는 경우 가능합니다.
본 항목은 본 항목으로 접근이 가능할 때 내회 전면 입국을 한
교집하는 경우 가능합니다.
본 항목은 본 항목으로 접근이 가능할 때 내회 전면 입국을 한
교집하는 경우 가능합니다.

의학연구심의위원회위원회
고려대학교안암병원
KOREA UNIVERSITY ANAM HOSPITAL



챗봇

전문성

글로벌

근거기반

朝鮮日報

IT Chosun

实时新闻 기업 모빌리티 게임·라이프 메타버스·AI 파이낸스 블록체인 사람 칼럼 테크잼

게임·라이프 > 메타버스·AI

“정신건강 쟁진다”…AI 챗봇 심심이, 우울증 권위자 함병주 교수 영입

변인호 기자

입력 2023.05.10 09:56

AI 챗봇 ‘심심이’가 정신건강을 쟁겨주는 디지털 헬스케어 서비스로 진화한다. 이는 심심이주식회사가 21년 동안 심심이를 서비스하면서 다양한 이용자가 가벼운 스트레스부터 우울과 불면, 산후우울증 등의 고민을 심심이에 털어놓는다는 점에서 착안됐다.

NEWSIS

AI챗봇 심심이, 더 똑똑해졌다…챗GPT활용 서비스 출시

카사등록: 2023/03/06 15:07:42
최종수정: 2023/03/06 15:16:47

챗GPT 공식 응용 프로그램 인터페이스 활용
GPT-3.5 터보 모델, 모바일 웹페이지로 구현

ChatGOV®

사용자: 수출 기반 국가에서 무역수지가 지속적으로 적자일 때 어떤 정책이 필요할까
ChatGOV:
1. 수출 다각화 정책: 수출 산업의 다각화를 촉진하는 것으로, 단일 수출 대상국에 의존하지 않고 다양한 시장을 개척해 수출 증가를 이루는 것이 중요하다.
2. 수입 대체 정책: 국내에서 생산 가능한 제품을 국내 시장에서 판매하거나 국내 산업에서 필요로 하는 원자료와 부품의 생산성을 촉진하여 수입 종속도를 낮추는 것이 중요하다.
3. 외국인 투자 유치 정책: 외국인 투자를 유치해서 수출 산업 확장 및 생산 기반 강화를 촉진하는 것이 필요하다.
4. 수출 중심 경제에서 내수시장 발전으로의 전환 정책: 외부 수요에 미온하는 수출 산업 대신 국내 시장을 중심으로 경제기반을 만들어 내수시장을 확대하는 것이 중요하다.

MTN 뉴스

타트업, '챗GPT' 들어올 때 노저어라…K 챗봇 시장 선점 누

네이버 뉴스 운석진 기자
2023/03/16 17:41:44

SimSimi

MTN 챗GPT 들어올 때 노젓는 韓스타트업



감사합니다.

ChatGPT 문장 생성을 활용한 한국어 감정 데이터 증강

이상협 · 이기성

오클랜드대학교, 뉴질랜드 · 중앙대학교

Augmented Stories: Korean Emotion Data using the ChatGPT text generation

증강된 이야기: CHATGPT 문장
생성을 활용한 한국어 감정 데이터

이상협 오클랜드 대학교

이기성 중앙대학교

증강된 이야기

1. 인간에 대한 허구적 이야기의 매력
 1. 심리적 요인: 일탈감, 상상력, 창의력, 감정 조절, 카타르시스
 2. 진화적 요인: 사회적 학습, 시뮬레이션, 응집력, 문화 전승
 3. 현실 적용: 공감, 시점 취하기, 인지 유연성, 스트레스 해소, 정신 건강
2. 인공지능 모델에 대한 허구적 이야기의 이점
 1. 감정 인식 학습 개선: 인간의 감정, 사회적 상호작용, 사고 과정 이해
 2. 풍부한 데이터셋: 다양한 감정 표현, 복잡한 시나리오, 장르, 문화, 역사
 3. 감성 지능 향상: 감정 조절, 카타르시스, 인지 유연성, 시점 취하기 적용

자연어 처리 모델의 문제

NLP 모델 성능: 학습
데이터 품질 및 양이
중요

데이터 부족 문제:
한국어 감정
단어사전에만 의존성

해결책: ChatGPT를
활용한 다양한 감정
예문 생성 및 증강 가능

ChatGPT 데이터 증강의 장점

제한된 텍스트
데이터셋 극복

다양한 NLP
작업 대안
제공

대용량 학습
데이터 생성

모델 정확도
향상

한정된 리소스에서의 효과적

인간 주석자 부담 감소

복잡한 전문 작업 집중 가능

다양한 도메인 및
작업에 적용 가능

ChatGPT를 활용한
데이터 증강: 기준 방식
한계 극복

감정 감지 모델 학습에
다양한 예문 생성 가능

데이터 주석 분야 혁명
및 NLP 및 AI 분야 혁신
촉진

기술 발전: 모델 학습 및
데이터셋 구축 효율성
및 정확도 향상 기대

한국어 감정 데이터셋 필요성

한국어 텍스트 데이터에서 정확하게 감정 표현을 파악하기 위함

영어와 한국어 간 번역 한계 및 부정확성 극복

단어(예, 가끔스레)와 의미(예, 불쌍하고 가여운 데가 있게),
용례(예, 할머니는 부모 없는 아이를 가끔스레 여기셨다), 그리고
감정(예, 연민)

단어 (19,617개), 의미 (19,617개), 용례 (9,335개)로 구성된 한국어
감정 데이터셋

기준 데이터 증감 방법의 한계 (동의어 교체, 역번역, 무작위 삽입,
삭제, 교환)

ChatGPT를 활용한 데이터 증강 방법

- 높은 정확성과 다양성 제공
- 문맥과 용례 고려한 예문 생성
- 감정 표현 인식의 정확도 향상
- "'느닷없이'라는 단어를 이용하여 '놀람'의 감정을 내포하는 일상의 긴 문장 20개를 만들어줘"
- '어제는 느닷없이 강한 비가 내려서 전혀 예상치 못한 일이 벌어졌어요.'
- '아침에 일어나자마자 느닷없이 전원이 나가버려서 허우적대던 그날의 기억이 있어요.'
- '지난 주말에는 느닷없이 친구가 찾아와서 당황하고 그랬어요.'

기대 효과



- 한국어 감정을 포착하는데 기존 모델이 부족함
- 한국어 감정 데이터셋 및 효과적인 증강 기술 필요
- 한국어 감정 분류 모델 성능 개선 및 한국 문화와 언어에 적응



언어 모델의 발전

- BERT (2018): Google의 혁신적인 NLP 기술, 맥락 인식 및 다양한 기능 제공
- RoBERTa (2019): BERT보다 성능 향상
- XLM-RoBERTa-Large
- RoBERTa 모델과 XLM (다국어 언어 모델) 결합
- 2.5TB의 방대한 데이터셋과 다양한 웹 텍스트, 위키백과 기사로 훈련

XLM- RoBERTa- Large의 장점

- 다양한 언어 지원: 한국어를 포함한 100여 가지 언어
- 언어 간 학습: 한 언어에서 배운 지식을 다른 언어에 적용, 적은 데이터로도 높은 성능 도달
- 높은 성능: 최신 RoBERTa 기술 기반
- 강인한 일반화: 다양한 언어 상황에서 신뢰할 수 있는 성능 제공
- 감정 분류에서 XLM-RoBERTa-Large 적용
- 한국어 감정 분류에서 높은 수준의 결과 예상
- 연구의 신뢰성과 효율성 향상



한국어 감정 감지 모델들 평가절차

- 첫번째 훈련 과정 세 단계
- 단어 기반 훈련 (39,234개)
- 단어와 의미 기반 훈련
- 단어, 의미, 및 ChatGPT 생성 문장 기반 훈련
- ChatGPT 생성 문장을 400,166개 적용



데이터셋 규모와 균형 고려

- 데이터 확장 및 리샘플링 기법 사용
- 균형 감정 모델 훈련
- 24개 감정 별 2만개씩 총 ChatGPT 생성 문장을 480,000 개 적용

감정 범주 축소 Ekman 6개 감정

Ekman (1972)의 6개
기본 감정과 중립
감정으로 통합

네 가지 모델 다시
훈련하여 성능 변화
관찰

통합의 이점

간소화 및 이해력
향상

국제적 비교 가능성

분석 및 모델링
간소화

학문적 유익성 및
활용도 향상

평가 지표 - F1 점수

- 주요 평가 지표 - F1 점수
- 정밀도와 재현율의 조화평균
- 감정 분류 문제에 적합한 평가 지표
- 입력 변수: 단어 벡터 토큰, 세그먼트 임베딩, 위치 임베딩
- 출력 변수: 24개의 한국어 감정 및 6개의 Ekman 감정
- 모델 평가: 정밀도, 재현율, F1-점수, 지지도

첫번째 연구, 24가지 감정 단어 모델 결과

- 첫 번째 모델 결과 (단어 기반)
- 전체 정확도: 0.41 (41% 정확도)
- 감정 별 가장 높은 F1-점수: 혐오 (0.48), 분노 (0.50), 슬픔 (0.53)
- 낮은 성능: 중립 (0.00), 질투 (0.00), 설렘 (0.01)



- 두 번째 모델 결과 (단어와 의미 기반)
- 전체 정확도: 0.46 (46% 정확도)
- 가장 높은 F1-점수: 질투 (0.58), 행복 (0.56), 수치 (0.43)
- 낮은 성능: 중립 (0.17), 연민 (0.25), 공포 (0.36)

첫번째 연구, 단어+의미 모델 결과



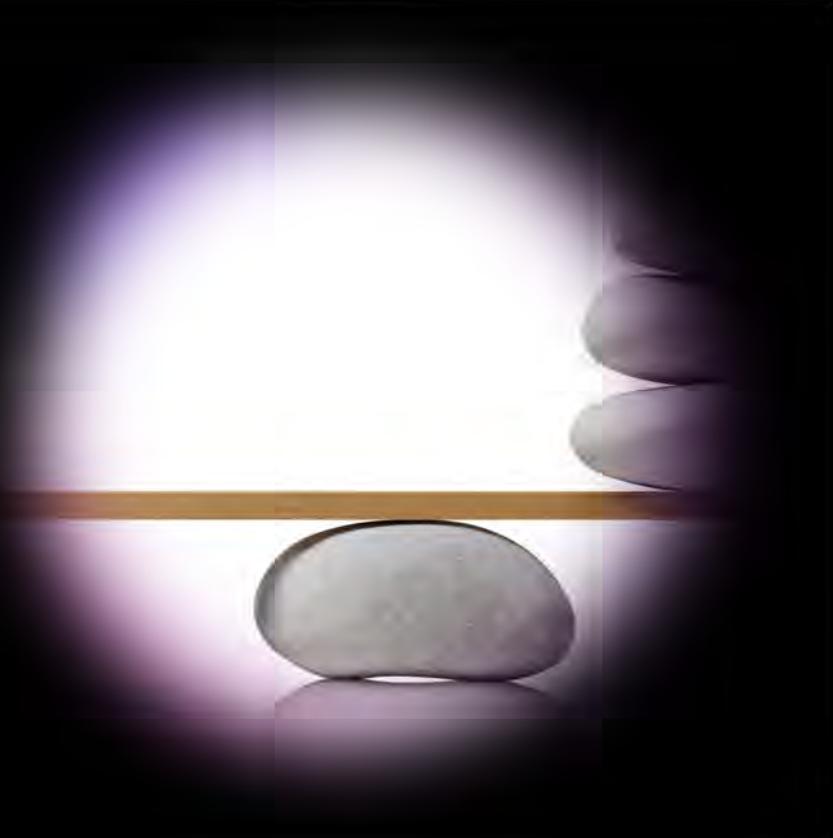
첫번째 연구, 단어+의미+ ChatGPT 모델 결과

- 세 번째 모델 결과 (단어, 의미, ChatGPT 문장 기반)
- 전체 정확도: 0.73 (73% 정확도)
- 가장 높은 F1-점수: 행복 (0.79),
분노 (0.80), 재미 (0.80)
- 낮은 성능: 중립 (0.50), 공포 (0.61)



두 번째 연구, 데이터 샘플링

- 데이터 샘플링을 통한 일부 감정 카테고리 정확도 개선
- 균형 잡힌 데이터셋 구축 (24개 감정별 각각 2만 개)
- 균형 감정 모델 정확도: 75%



세 번째 연구, Ekman의 6가지 감정 통합

- 세 번째 모델 정확도: 81% (높은 성능 향상)
- 균형 에크만 감정 모델 정확도: 82% (가장 높은 정확도)



결론 및 기대효과

- 다양한 언어와 도메인에 대한 데이터 증강 방법 적용 가능
- NLP 작업 및 응용 분야에서 혁신과 발전 기대
- 데이터 주석 분야의 혁명적 변화를 가져오며, 인공지능 분야에서 응용과 혁신 촉진



감사의 말



이번 연구를 진행하며
도움을 주신 모든 분들께
진심으로 감사드립니다.



본 연구의 결과가 감정 인식
및 인공지능 분야의 발전에
도움이 되기를 바랍니다.

제6회 인공지능인문학 국제학술대회

세션 2

심포지엄: 인공지능과 기계의 창의적 오류

사회: 박평종 (중앙대학교)

토론: 박주석 (명지대학교)

박충식 (유원대학교)

이미지 생성 인공지능의 미학적 가능성

박평종

중앙대학교

이미지 생성 인공지능 의 미학적 가능성

<미증유의 얼굴> 개요

- 이미지 생성 인공지능의 예술적 가능성
- 알고리즘의 '오류'로 산출된 이미지의 시각적 효과 분류
- 생성 이미지의 미학적 가능성 타진, 미술사/사진사에서 지니는 의미 진단

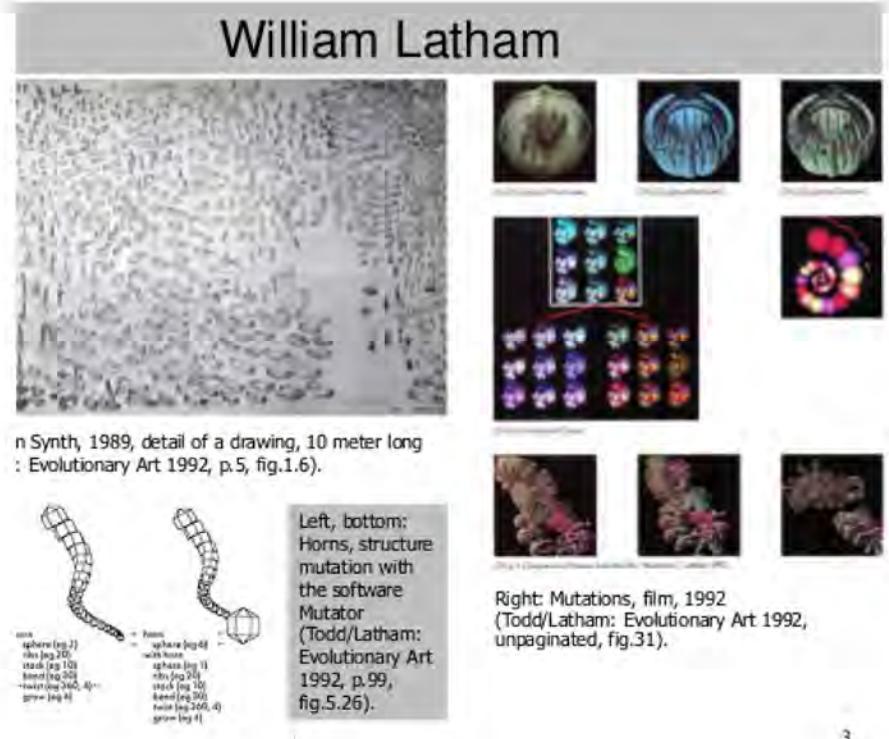


컴퓨터의 창의성: 마가렛 보덴

- 창의성: "새롭고 가치 있는 사고를 산출하는 능력"
- 역사적 관점: 과거에 한번도 존재하지 않았어야 함
- 심리적 관점: 새롭게 인식되어야 함(역사적 관점을 포괄)
- 조합적 창의성(Combinational Creativity)
- 탐색적 창의성(Exploratory Creativity)
- 변형적 창의성(Transformational Creativity)

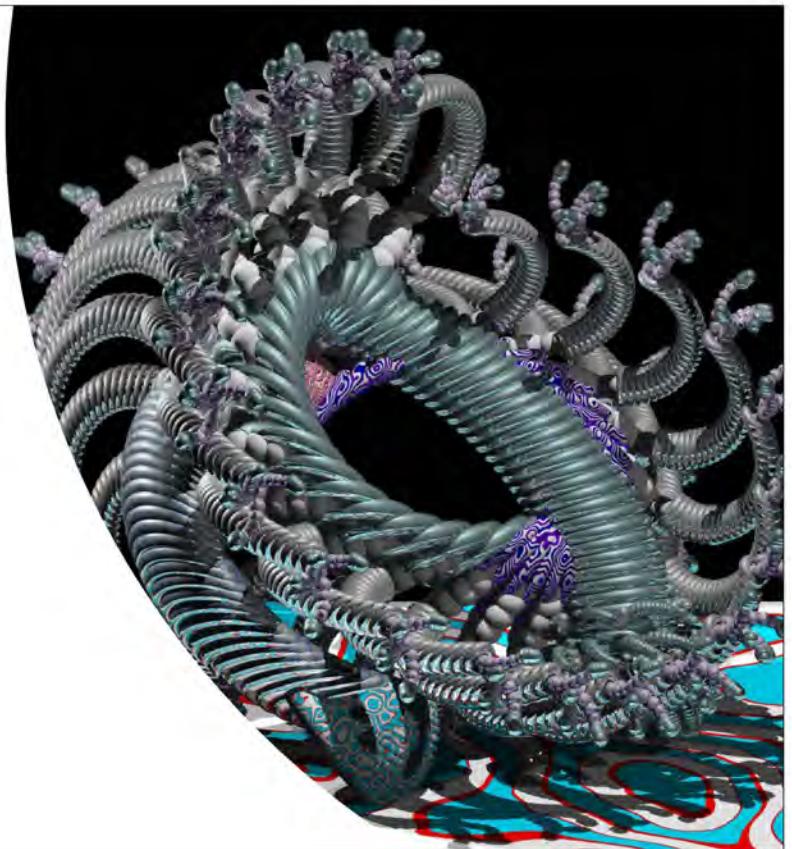
조합

- 익숙한 사고를 낯설게 조합(unfamiliar combination)하는 방법
- 예 : 콜라주, 몽타주
- 래섬의 인공 유기체 : 단순한 기본 형태들이 서로 착종하면서, 즉 조합을 통해 새로운 형태를 만들어내기 때문



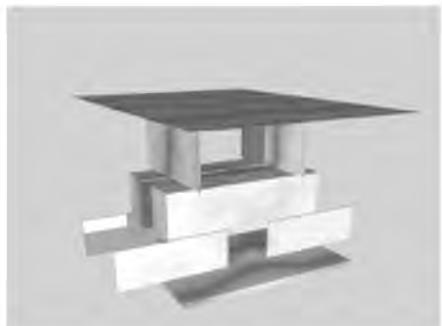
래섬(William Latham)의 인공유기체

- 진화하는 형태(Evolving form) : 구와 원뿔, 원환과 같은 단순한 형태들이 규칙에 따라 단계적으로 착종하면서 변화
- 기본 형태는 유전자처럼 작용, 이종교배나 변형을 통해 진화
- 작가의 개입 : 자연스러워 보이는 형태의 선택



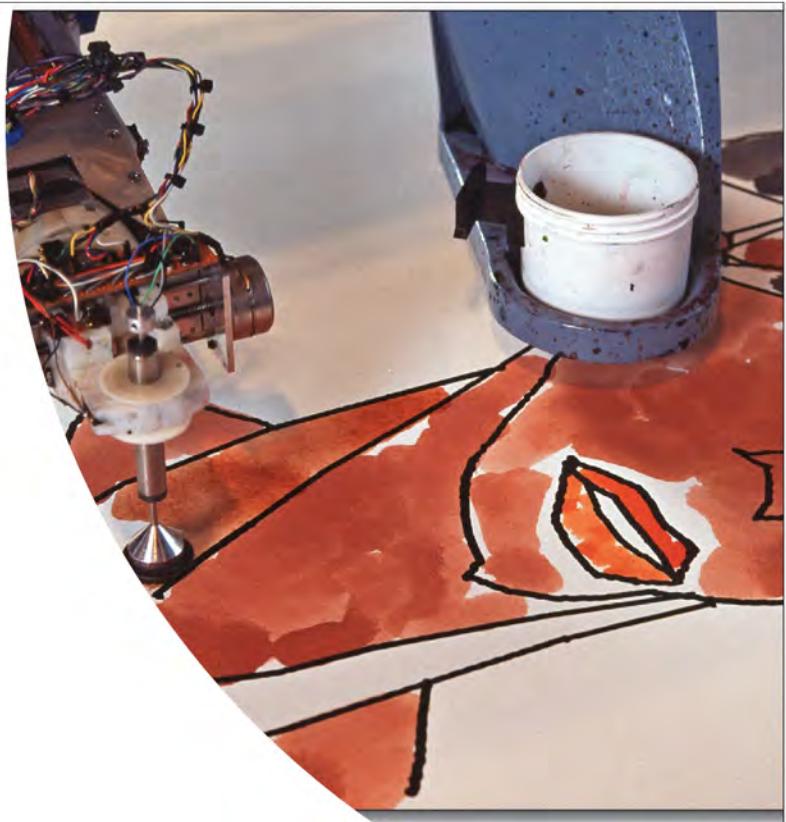
Peter Bentley의 진화 알고리즘

- 일반 진화 디자인 시스템(Generic Evolutionary Design system) : 테이블의 구성요소들을 복잡한 배치를 통해 서로 결합시켜 형태를 발전시킴
- 적합도 함수 적용시 테이블의 기본원칙을 고려



탐색(exploration)

- 문화적으로 용인된 사고에 의지하여 새로운 생성규칙(generative rule)을 찾아나가는 방식
- 한계 : 예컨대 문학이나 언어학에서 새롭게 단어를 배열하고자 할 때 새로운 생성규칙은 기본 문법을 준수해야 함
- 만약 이를 무시한다면 '새로운' 생성규칙은 결코 수용될 수 없음
- 예 : 헤럴드 코헨의 아론(Aaron)

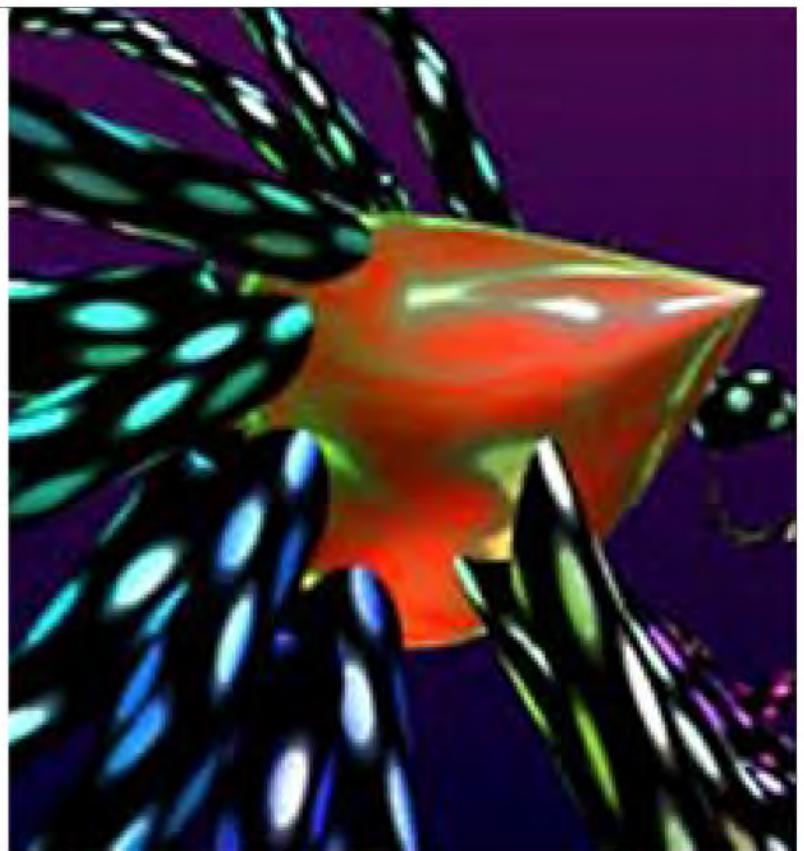


AI Painter AAaron

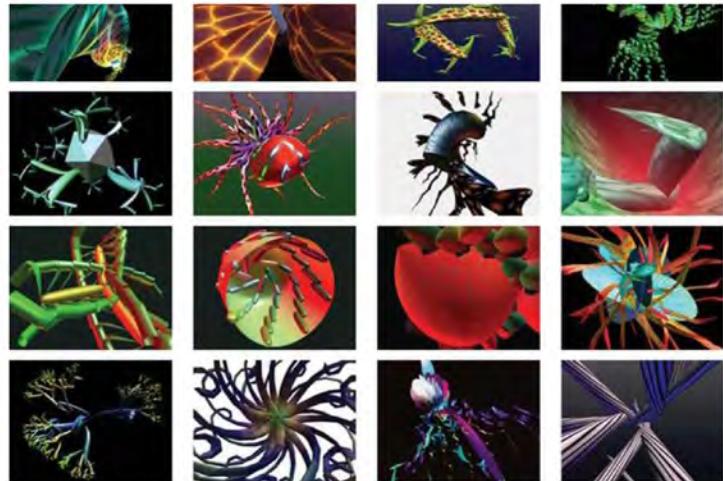


변형(transformation)

- 사고나 형태를 결정하는 범위에 변화를 주어 차이를 만들어 내는 방식
- 가장 급진적 : 이 방식으로 산출된 사고(정보, 형태)는 새로울 뿐만 아니라 이전에 생산됐던 것들과 근본적으로 다르기 때문
- 예 : 칼 심스의 <Galapagos>



갈라파고스

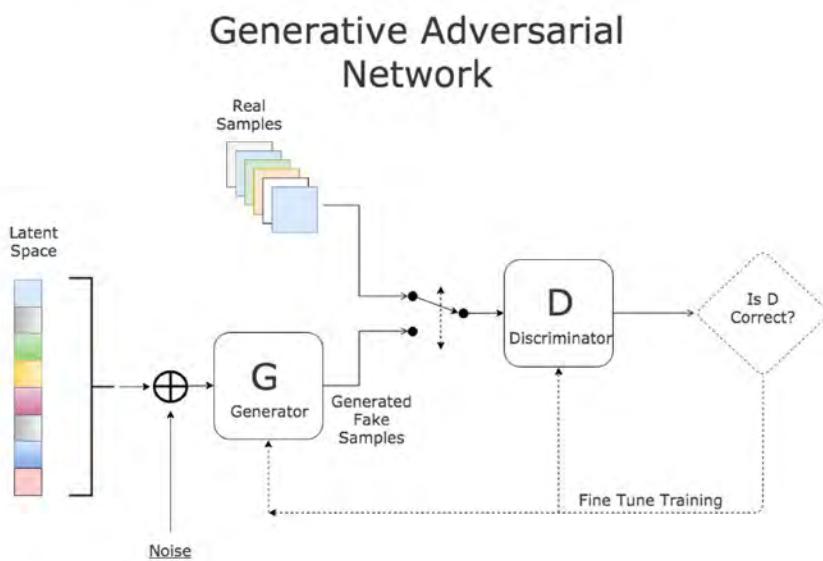


연구대상

- 이미지 생성모델을 적용한 인터넷 플랫폼
- Thispersondoesnotexist.com: StyleGAN/StyleGAN2를 적용
- 대부분의 이미지는 실제 인물과 흡사한 '사진같은' 얼굴
- 그러나 오류가 발생 : '놀랍고 탁월한' 왜곡 이미지
- 패턴과 유형 분류



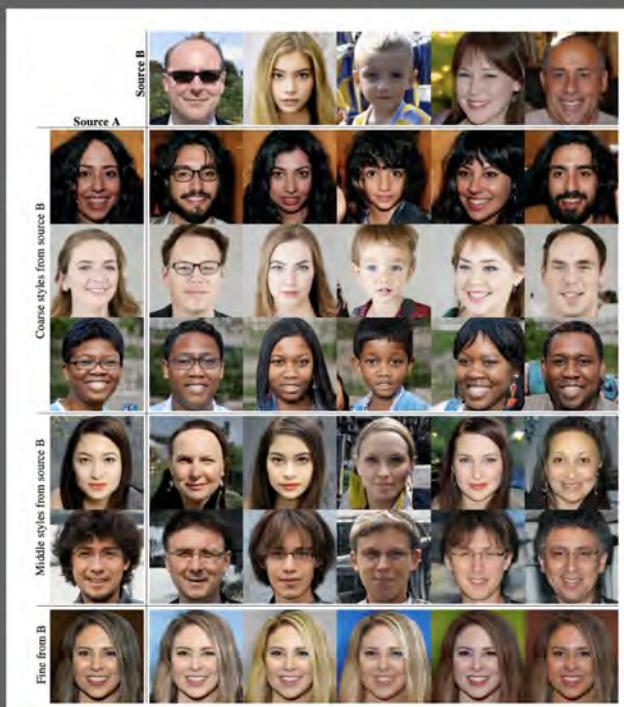
생성 모델: GAN(Generative Adversarial Network)



- 생성자(Generator): 실제 데이터와 가까운 가짜 데이터의 생성이 목표, 위조지폐법
- 판별자(Discriminator): 진짜 데이터와 가짜 데이터를 정확히 구분하는 것의 목표, 경찰
- 적대적(협력적) 경쟁을 통해 서로의 능력을 극대화시키면서 상호 발전

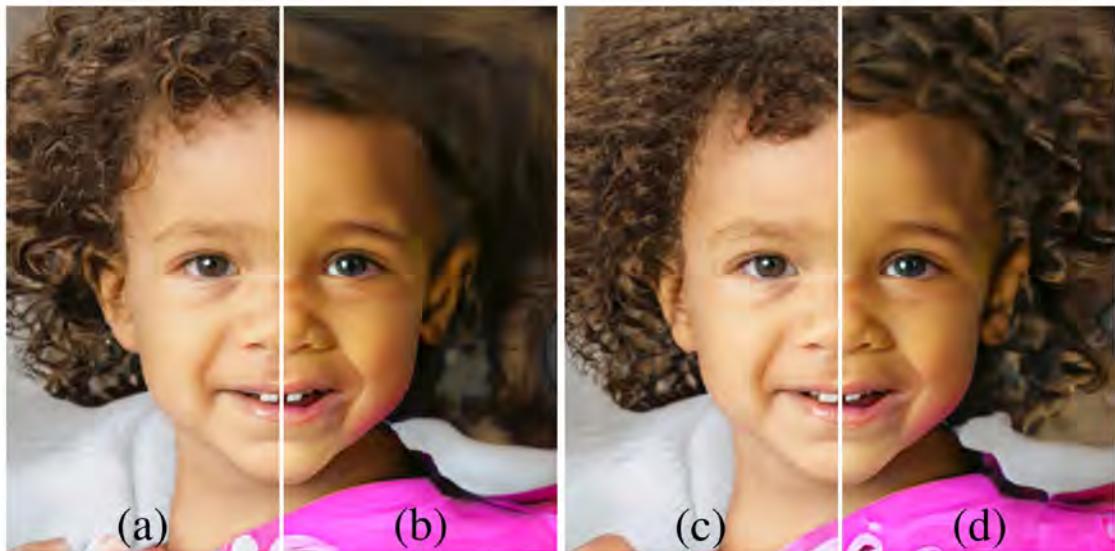
이미지 스타일 변환: StyleGAN

- 각 레이어에서 해당 이미지의 스타일을 조정
- 저해상도(coarse resolution, 4x4-8x8) : 성별, 포즈, 얼굴 형태, 일반적인 헤어스타일
- 중간 해상도(middle resolution, 16x16-32x32) : 얼굴의 부분적 특성, 디테일한 헤어스타일, 입의 개폐
- 고해상도(fine resolution, 64x64-1024x1024) : 눈동자 색, 머리카락 색, 조명에 따른 얼굴색의 변화, 얼굴의 미시적 특성



StyleGAN의 합성

노이즈 효과 : a(모든 레이어에 노이즈 추가), b(노이즈 x),
c(고해상도에만 노이즈), d(저해상도에만 노이즈)



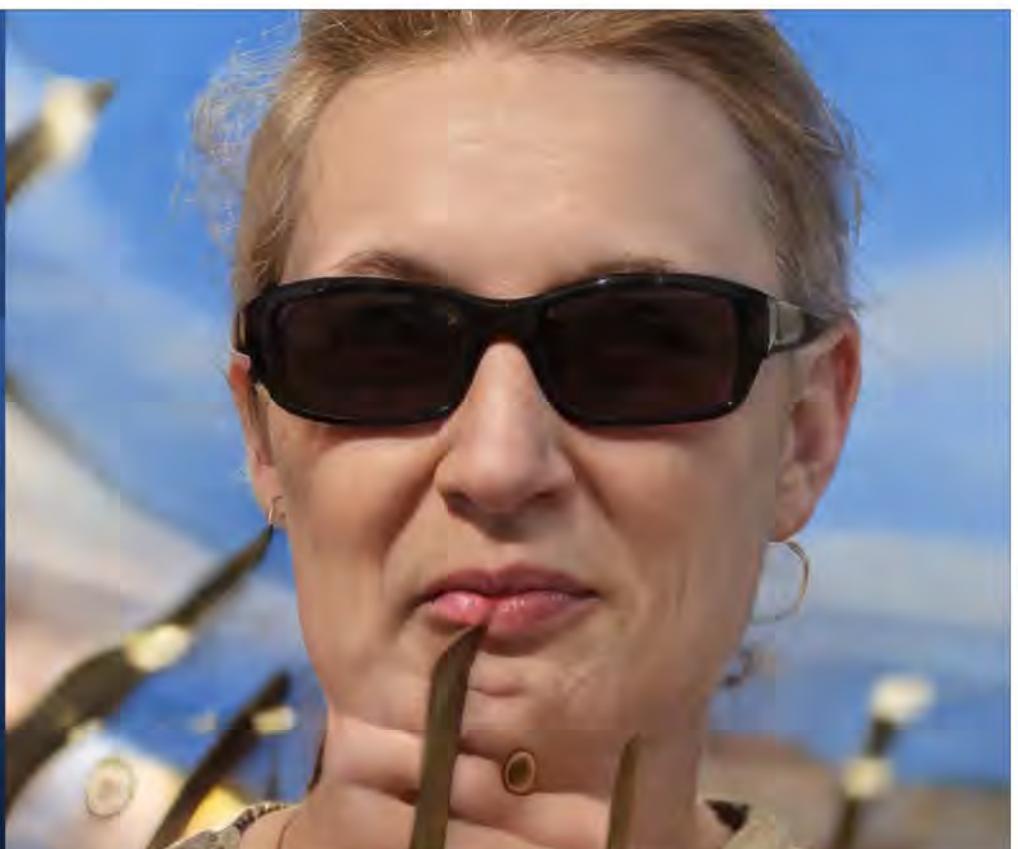
StyleGAN의 문제

- 1. 물방울 얼룩: 시각적으로 나타날 뿐 아니라 생성자 네트워크 내부에서도 발견, 64x64 해상도에서 시작하여 사라지지 않음, 시스템 문제
- 2. 얼굴 전체와 부분적 요소들의 불일치: 정면이 이동해도 부분(눈, 코, 치아 등)이 본래 위치를 고수

StyleGAN2

- 1. 물방울 얼룩의 제거: 스타일 합성을 수행하는 AdaIN(Adaptive Instance Nominalization)의 표준화 방식을 개선, 실제 통계치의 표준화(작은 노이즈가 표준화 과정에서 부풀려짐)에서 추정 통계치의 표준화로
- 2. 전체와 부분의 불일치 해결: Progressive Growing 방식 개선 (StyleGAN이 저해상도에서 고해상도 학습으로 진행하는 과정에서 각 레이어는 분리되어 학습됨)

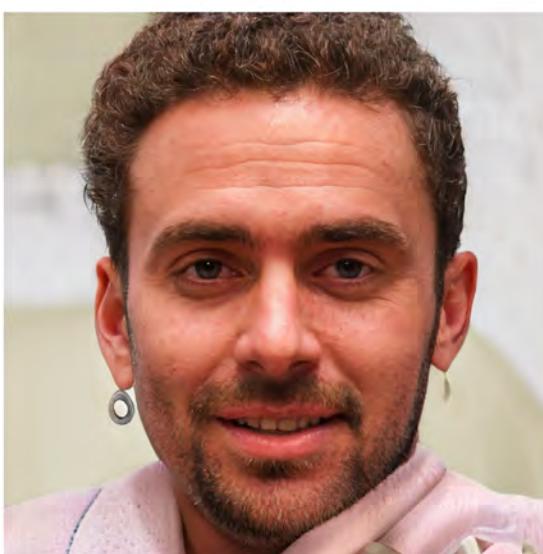
디테일의
파격: 입술의
포스



민감한 피부/이마에 달팽이



자석 귀/제 3의 눈



잘린 볼/봉합의 흔적



변형된 신체:
달팽이 손

각진 목/손가락을 뜯어먹는 남자



구겨진 입(1)/구겨진 입(2)



늘어난 손(1)/늘어난 손(2)

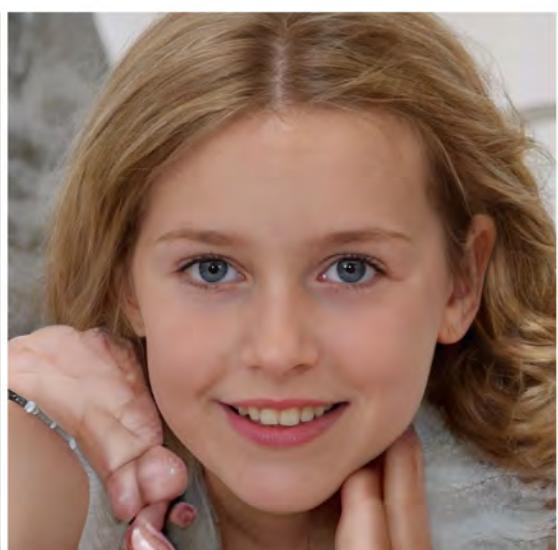


손가락
분신술

입술에 달라붙은 손가락/이마에 달라붙은 손가락

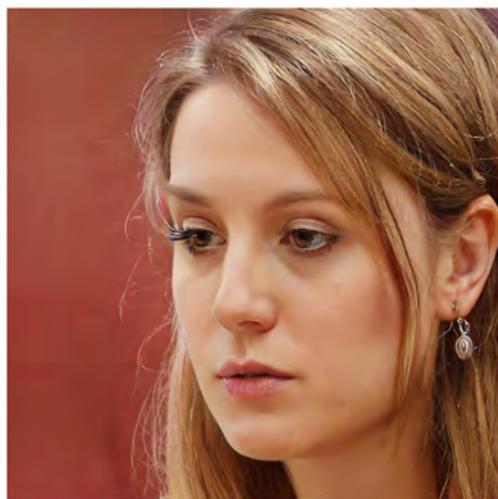


입을 쥐어뜯는 여자/조각난 손가락





궤도이탈: 기막힌 8자



눈을 파고든 머리칼/귀에서 자라난 머리칼

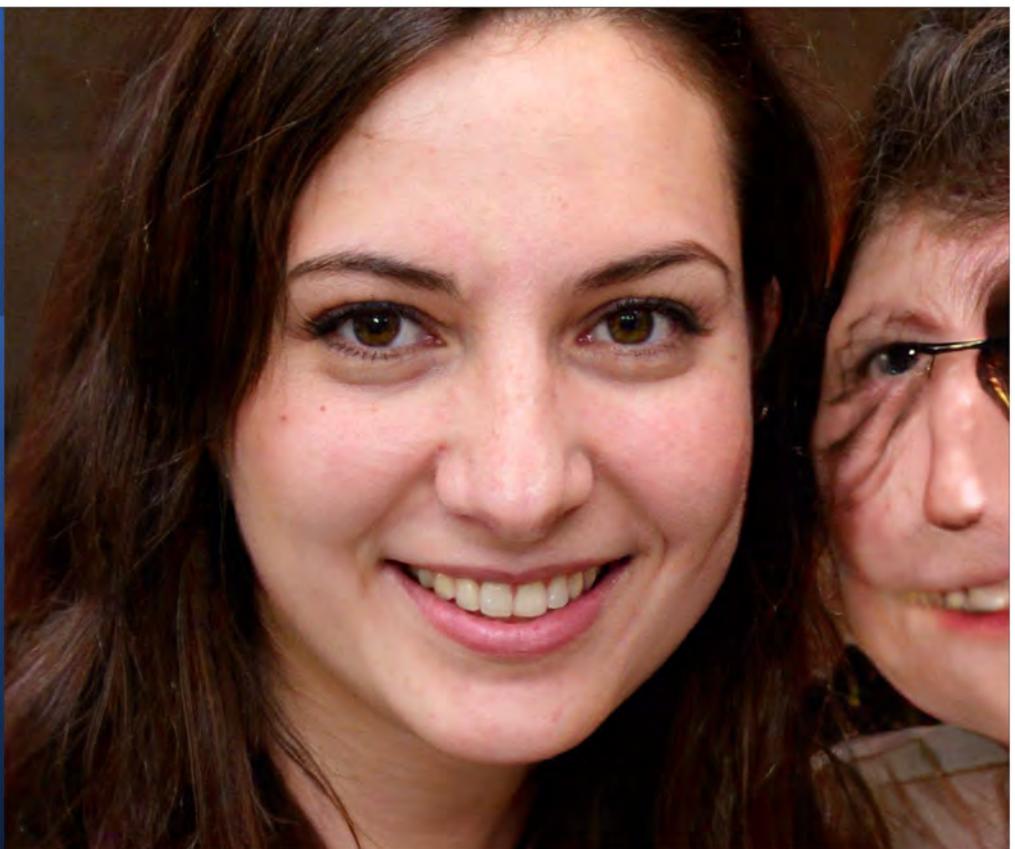
끊어진 머리칼/손가락 빗질



어지럽게 널린 머리칼/조각난 피부



괴물친구들



괴물친구들(2,3)



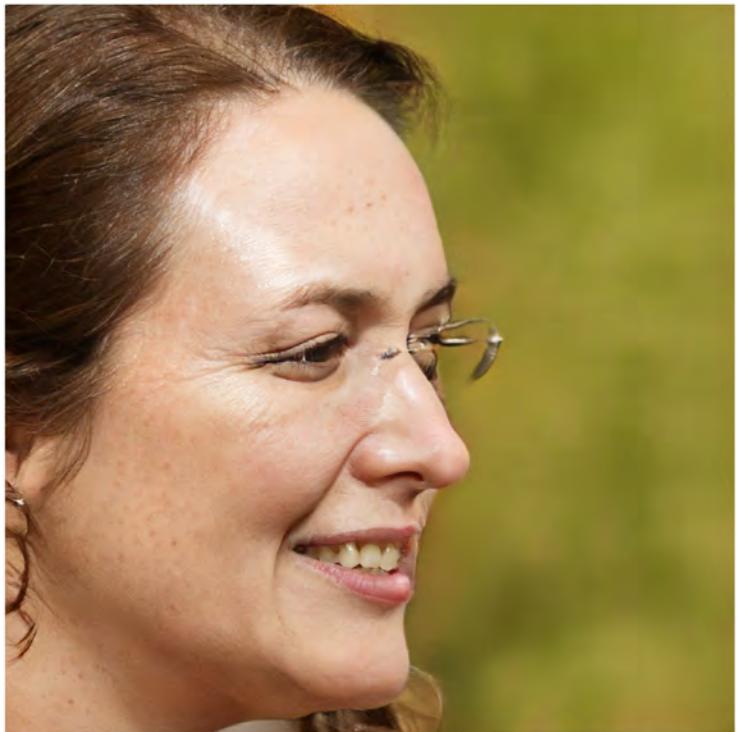
괴물친구들(4,5)



괴물친구들(6,7)



사라진
안경



사라진 안경(2,3)



사라진 안경(4,5)



얼룩의 진화:
카니발의
여신

상처투성이 얼굴



금박머리 여인/화려한 머리장식



메타모르포즈: 다프네의 후예



물질의 친화력



불가사의한 힘



빨간머리 맨/사라진 모자



컬러의 도발:
형광머리카락의
여인



그린핑크 아이쉐도우/블루가 좋아



초현실적 초록머리/화려한 헤어스타일



초록 유전자





앰비발란스:
턱수염이 난
여인

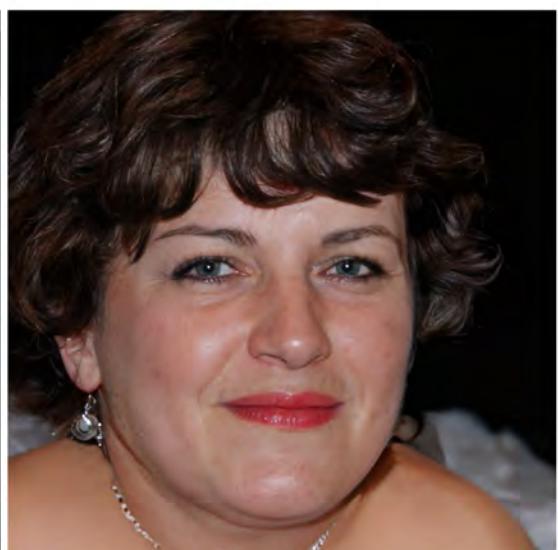
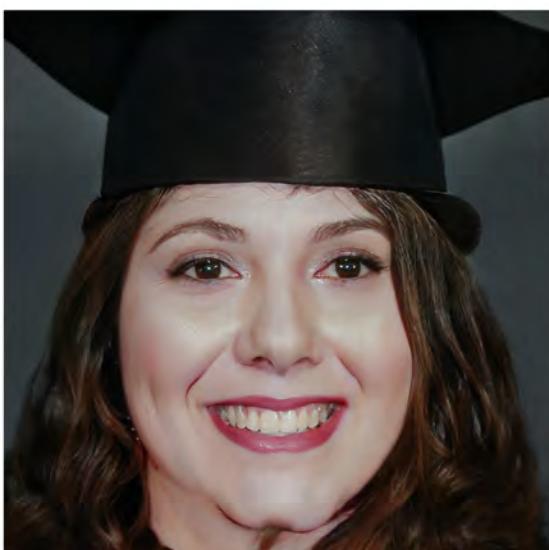
과도한 곱슬머리/가녀린 입술



루즈 칠한 희머리 소년/미세모에 덮인 여인



머리칼이 턱에 붙은 여인/트랜스젠더



돌연변이: 턱이 갈라진 남자/게으른 남자



깔때기 얼굴/볼이 찢어진 남자





놀라운 능력

신비한 초록눈



신비한 초록눈



착란의 징후:
눈꺼풀을
뜯어내는
사람

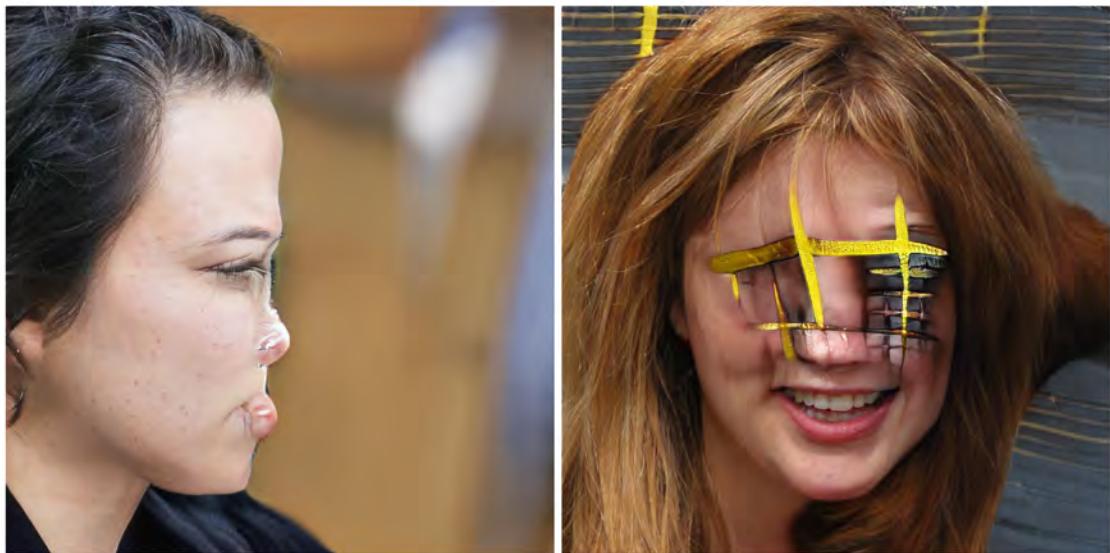
머리칼 폭풍/기구한 운명



부서진 철가면/다다의 문법



혹독한 환원/소극적 반달리즘



미술사, 사진사 속의 왜곡



Hans Holbein,
The
Ambassadors,
1533



세부



왜곡

살바도르
달리, The
Persistence
of memory,
1931



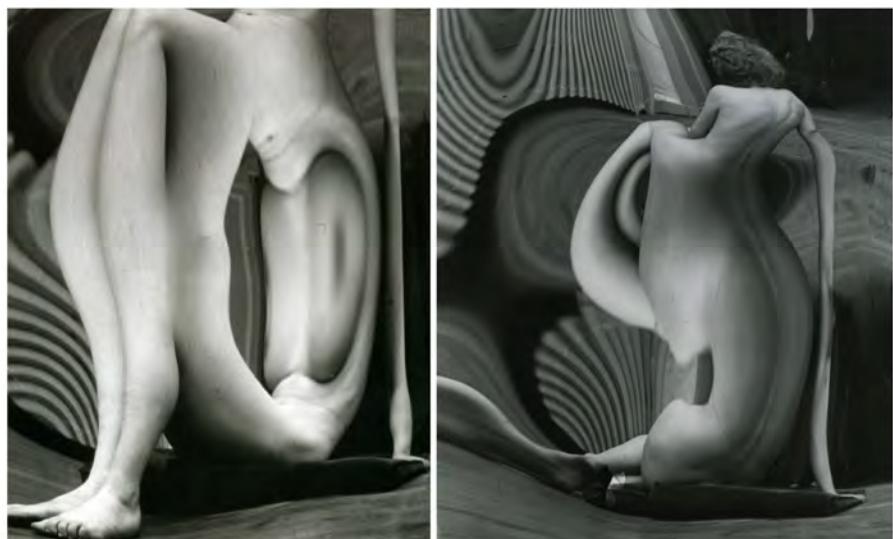
살바도르 달리,
(스페인)내란의
예감, 1936





앙드레 캐르테즈, Distortion

앙드레
캐르테즈



앙드레
케르테즈



Aziz +
Cucher,
Dystopia,
1994-95







결론

- 실재의 왜곡과 변형, 우연을 통한 이질적 요소들의 비논리적 제휴, 가상 세계는 이미지의 역사에서 항상 탐구의 대상이었음
- 제한된 시각경험을 풍부하게 해 줌
- 새로운 가시성을 제안함으로써 인간의 시각을 확장시켜 줌
- 생성 이미지의 미학적 가능성을 수용할 것인지 여부는 인간의 몫

에러와 사진예술

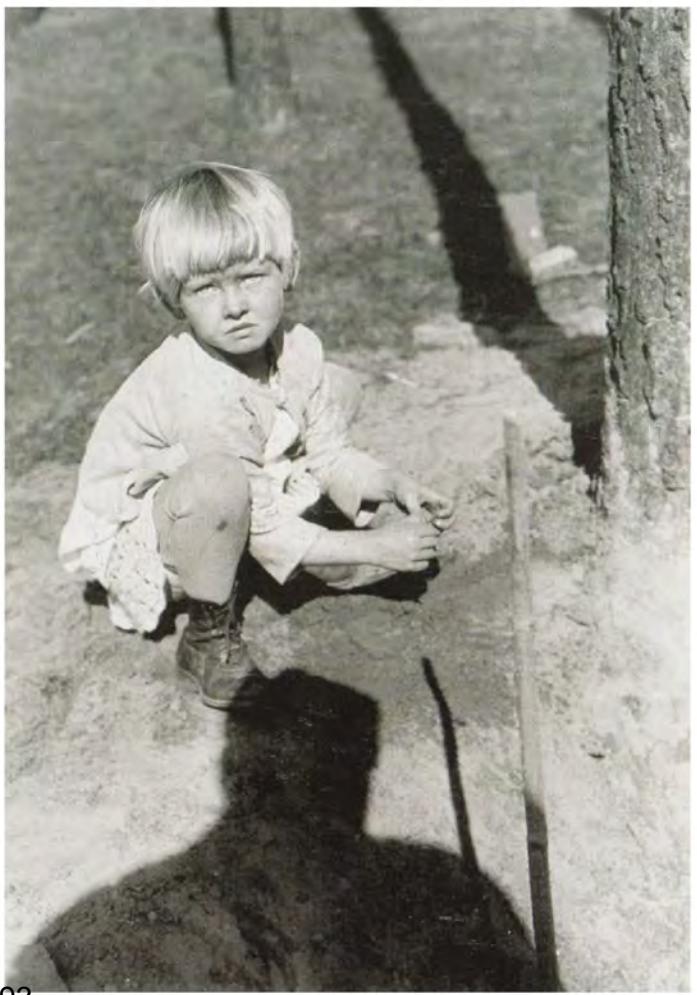
박상우

서울대학교

에러와 사진 예술

Error and Photography Art

박상우(서울대 미학과 교수)



Error 미학

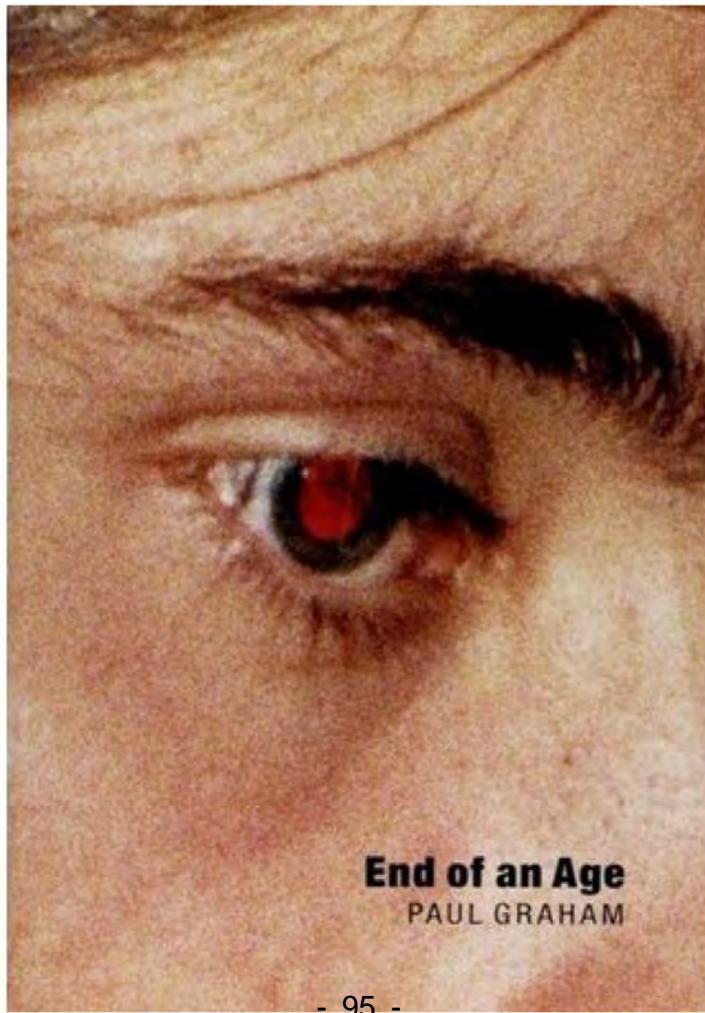


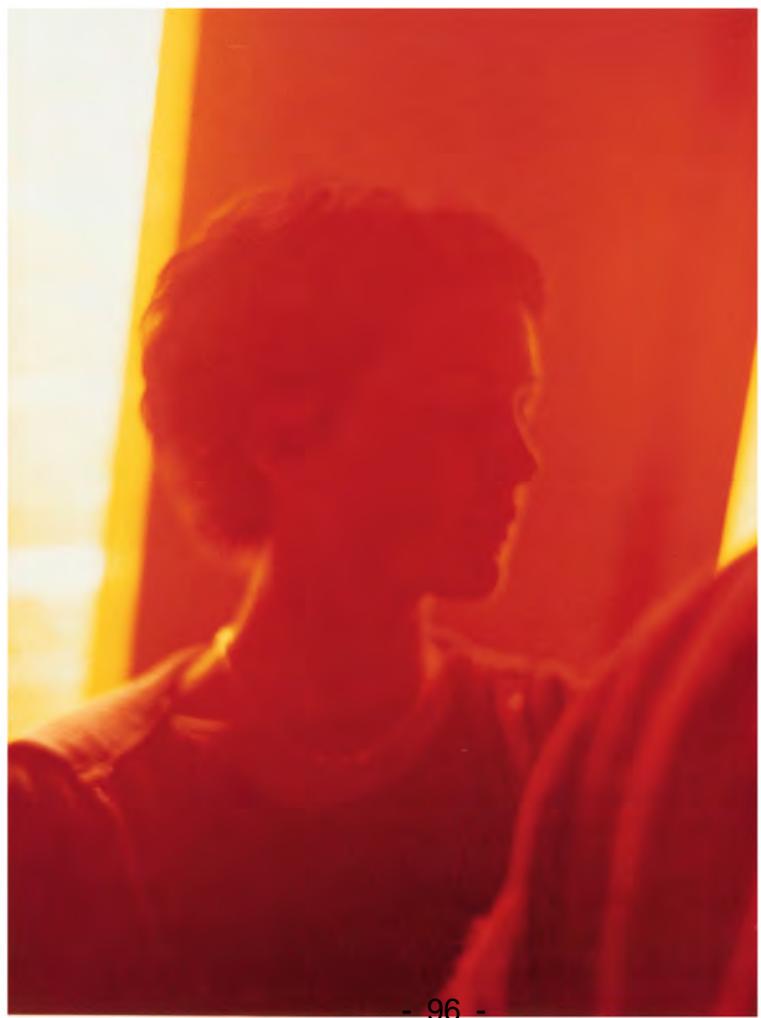
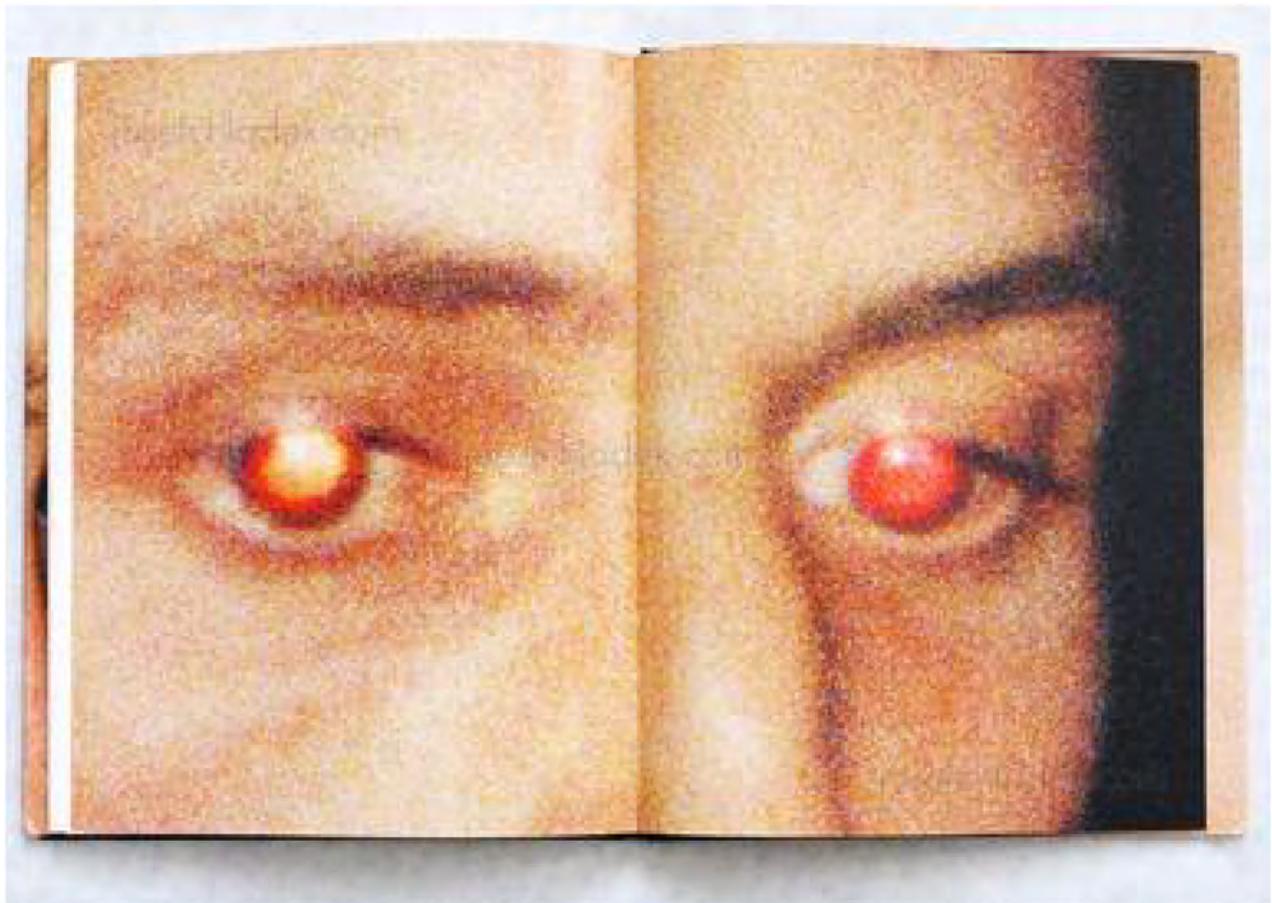
- “아마추어 사진가에서는 오류로, 프로에게는 완전한 실패로 간주된 사진이 예술가에게는 흥미로운 이미지로 인식될 수 있다.”
- 현대미술에서 일군의 작가들은 사진가들의 오류에 주목한다.”
- Clément Chéroux

perfect 미학



에드워드 웨斯顿
Charis Nude, 1936





LES SURPRISES DE LA PHOTOGRAPHIE, — par LUC.



Le modèle a bouché un tout petit peu.



Ce coup-ci, c'est l'appareil qui a remué légèrement.



La plaque était mal préparée.



L'appareil était trop petit.



A peu près raté.



Deux clichés sur la même plaque.



Un ménage très uni (groupe).



Mauvaise mise au point.

실수/성공 판단 기준



- 전문가의 특정한 규칙
- "만약 고객이 사진을 원한다면 이 딱지를 간단히 떼어내면 됩니다."
- 유동적

영역에 따라 다름

- 아마, 프로, 예술 영역
- “[사진에서] 흐릿함, 잘린 머리 등이 무슨 상관이랴. 사진에 재현된 인물을 사랑한다면 유일하게 추억만이 중요할 뿐이다.”



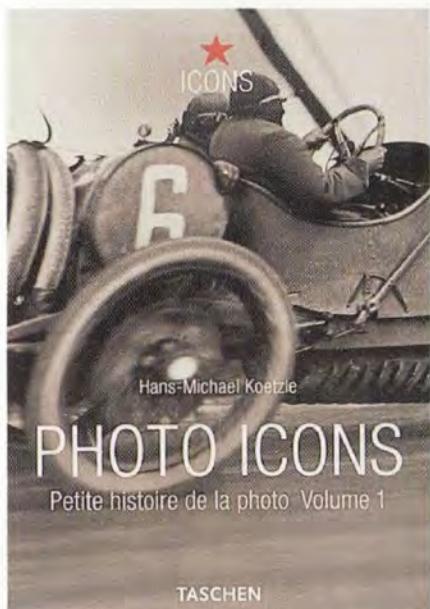
미술비평가 Laurent Boudier

시간에 따라 다름



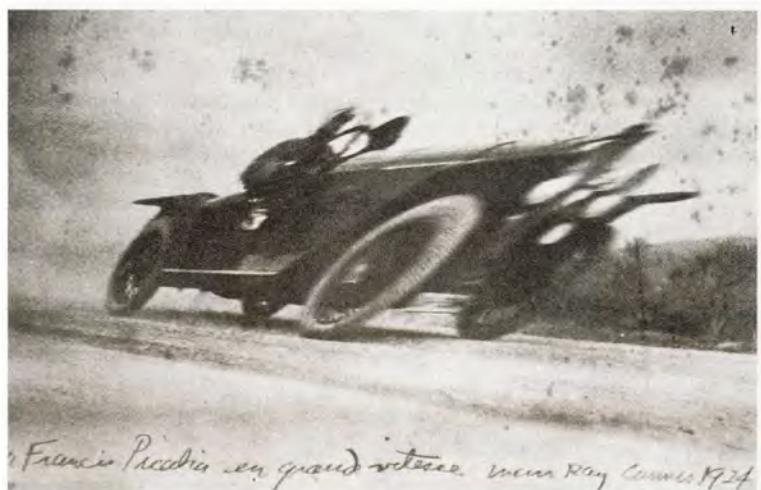
- “실패한 작품에 대한 판단은 그 작품을 평가하는 사람이 속한 시간에 따라 변한다.”
- Pierre Bayard (문학평론)

자크 앙리 라르티크. <자동차 경주>. 1913



- 40년후 라르티크 (1950년대)
- “실패는 매우 자연스러운 것이다.
. 실패는 좋은 교훈이다. 이 때문에 만족스럽지 못하다고 판단된 사진도 보관해야 한다.

왜냐하면 과거에 매우 만족하지 못한 것에서 3년, 5년 혹은 10년 안에 어떤 의미 있는 것을 발견할 수 있기 때문이다.”



에러사진 → AV 사진영향



“아마추어의 오류
사진들: 하이앵글,
로우앵글 같은 특
징들은 오늘날 우
리를 깜짝 놀라게
한다.”

(모흘리 나기)



그림자 오류

Nº 52 REVUE ILLUSTRÉE DE PHOTOGRAPHIE 29 Décembre 1901

Les PETITES MISÉRES DU PHOTOGRAPHE

L'Auto-Ombromanie

La photographie que nous reproduisons ci-dessous, et que nous devons à un de nos abonnés, n'est pas à proprement parler entachée de défectuosité accidentelle; elle devrait rentrer plutôt dans la catégorie de ces anomalies que nous avons appelées antrefois les *Surprises du Gélatino...*

La particularité qui nous engage à la placer sous les yeux de nos lecteurs, bien qu'elle ne présente par ailleurs aucun intérêt immédiat comme composition ou exécution, réside exclusivement dans l'ombre portée au premier plan par l'opérateur et par la personne placée à ses côtés.

Le soleil était fortement incliné sur l'horizon; ses rayons frappaient nos personnage sous une obliquité telle que leur ombre s'al-

houette les deux auteurs de la scène, devant acteurs à leur insu, et de cette constata on inférerait volontiers qu'il doit être fait de baser sur cette sorte d'auto-silhouette non prémedité, une nouvelle méthode d'ombromanie photographique.

Rien ne serait plus simple que d'imaginer un système d'écran inclinable, permettant de donner aux ombres registrées une absence de correction, consistant dans les déformations facultativement en gueule, en largeur diagonale, etc.

Par exemple, écran plan, disposé de façon à être perpendiculaire à la sectrice de l'angle formé par les rayons soleil et l'axe de l'objectif; conservant aux ombres des portions rigoureusement exactes.

Des mouvements d'inclinaison tendent à rapprocher de l'outil soit l'un des cadrans latéraux, soit l'une extrémité supérieure ou inférieure de l'écran, déformeraien silhouettes dans

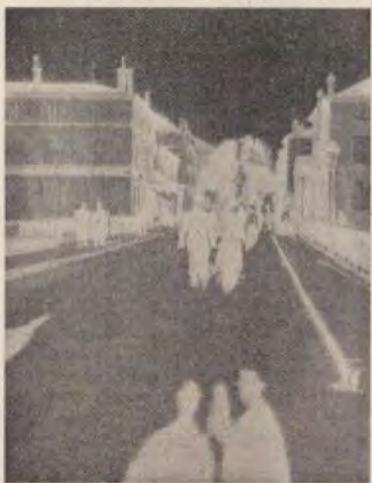


Fig. 1. — L'ombre de l'opérateur dans la composition.

<사진가의 작은 불행.

자신의 그림자>. 1901.



왜 그림자 사진 집착?



- 사진 투명성 신화 파괴
- 사진 본질 = Ombre-graphie

모흘리 나기. <아스코나>. 1926.

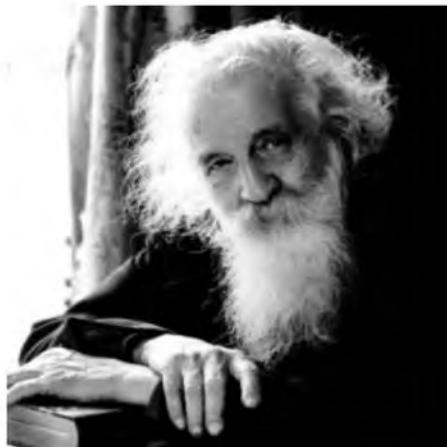


사진 근본 요소

- 빛
- 그림자
- Projection
- Reduction
- 테크닉
- 인간

에러의 힘

- 사진 매체의 주요 요소
드러냄
- “지식은 항상 그림자를
동반하는 빛이다.”



바슬라르



Lisette Model,
«Premier reflet», New York, 1939-1940



Lee Friedlander, New York, 1966





Ugo Mulas,
L'opération photographique.
«Autoportrait pour Lee
Friedlander», 2e vérification,
1970

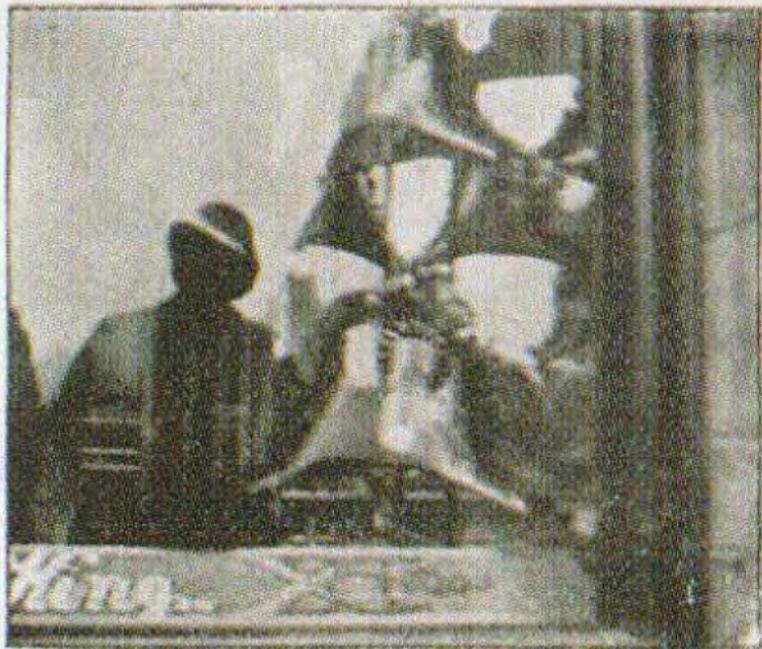


Denis Roche,
11 octobre 1987, Paris, la

reflection의 오류

Common Hand Camera Failures.

(Continued from last month.)



익명의 사진가. <휴대용 카메라의 흔한 실수>. 1902.



오젠 앗제. <봉 마르세>. 1926.
- 106 -





Man Ray *Fin de l'ère chrétienne*, 1925

N° 3 — Première année

15 Avril 1925

LA RÉVOLUTION SURRÉALISTE

1925 : FIN DE L'ÈRE CHRÉTIENNE

SOMMAIRE

L'Europe et l'Asie : Théodore Lessing	Glossaire : Michel Leiris.
Pamphlet contre Jérusalem : Robert Desnos.	L'amour des heures : Benjamin Péret,
Descriptions d'une révolte prochaine : Robert Desnos	Décadence de la vie : Jacques Baron.
La revendication du plaisir : Jacques Baron et Michel Leiris.	Un suppression de l'esclavage : Paul Éluard.
Rêves : Paul Éluard, Pierre Naville, Raymond Quencau,	L'activité du Bureau de Recherches : Antoine Artaud.
Illustrations : Giorgio de Chirico, Paul Klee, André Masson, Max Ray, Delé Sonbeam	Textes surréalistes : D. L., Max Morise.
ADRESSE AU PAPE, ADRESSE AU BRAHMANA LETTRE AUX MÉDECINS CHIENS DES ASILES DE FOIS	12 phrases de réveil : Maurice Béchet.
ABONNEMENT : les 12 Numéros : France : 45 francs Etranger : 55 francs	Beaux-arts : Pierre Naville.
Dépositaire général : Librairie GALLIMARD	LETTERE AUX ÉCOLES DU MONDE
11, Boulevard Raspail, 15 PARIS (VII)	LETTERE AUX RECEVEURS DES UNIVERSITÉS EUROPÉENNES

LE NUMÉRO :
France : 4 francs
Etranger : 5 francs



<초현실주의 혁명> 제 3호, 1925

- “반영은 더 이상 기술적 오류가 아니라, 미학적 명제들 중의 하나에 속하게 되었다.”



그것은 이때부터 아방가르드 사진의 시각적 준거로 인정받았다.”

Clément Chéroux

왜?

a) 전복

- “전복은, 특히 예술에서 작동할 때 새로운 힘들의 거대한 보고(寶庫)가 된다.”

• André Breton



b) 지각의 동시성

- “반영은 현대도시의 경험이 인간에게 유발하는 감각의 혼돈된 이미지”

Clément Chéroux,



반영 = 우연의 몽타주





Lisette Model,
«Premier reflet», New York, 1939-1940



Fauto-graphe



“내가 사진을 촬영할 때, 혹은 암실에 있을 때, 나는 일부러 모든 규칙을 따르지 않았다. 나는 결과물들을 가장 비상식적으로 섞어보았다. 나는 유효기간이 지난 필름을 사용했고, 화학과 사진에 영향을 줄 수 있는 최악의 것들만 행했다.”

Solarization



- “어느 날, 나는 암실에 있었다. 그리고 어두운 것에 갑갑함을 느꼈다. 그래서 램프를 켰다. 바로 이렇게 나는 솔라리제이션을 발견했다.”

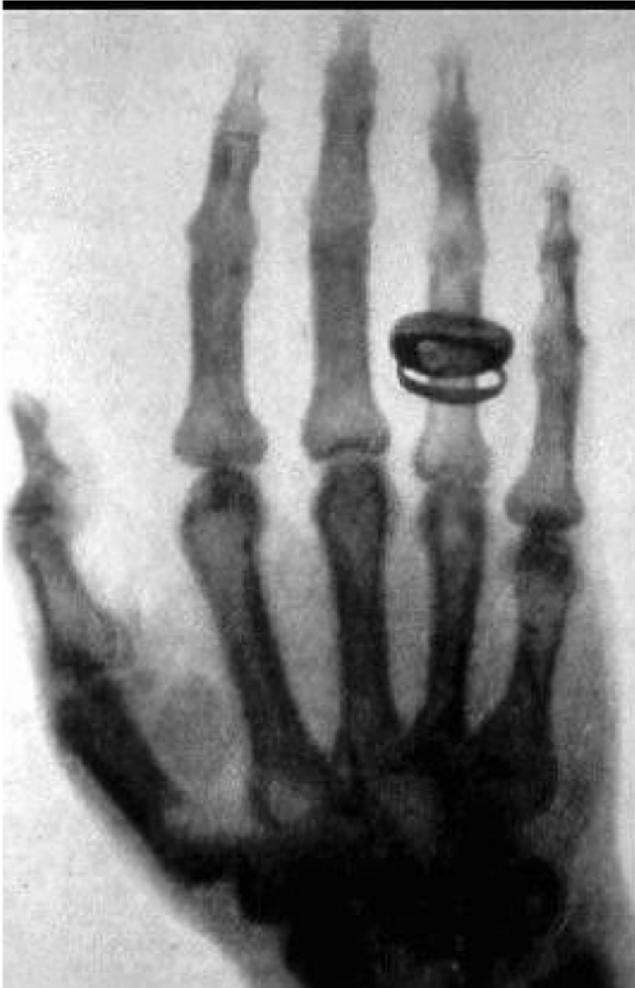
rayogram



“어느 날 암실에서 사진을 확대하고 있을 때, 하나의 오브제가 우연히 인화지 위에 놓이게 되어 흔적을 남겼다.”

serendipity

- 과학에서 실수와 우연이 지닌 생산과 창조의 힘



- X-RAY 발견
- 플레밍의 페니실린
- 방사선
- 전자기장

우연 + 통찰력

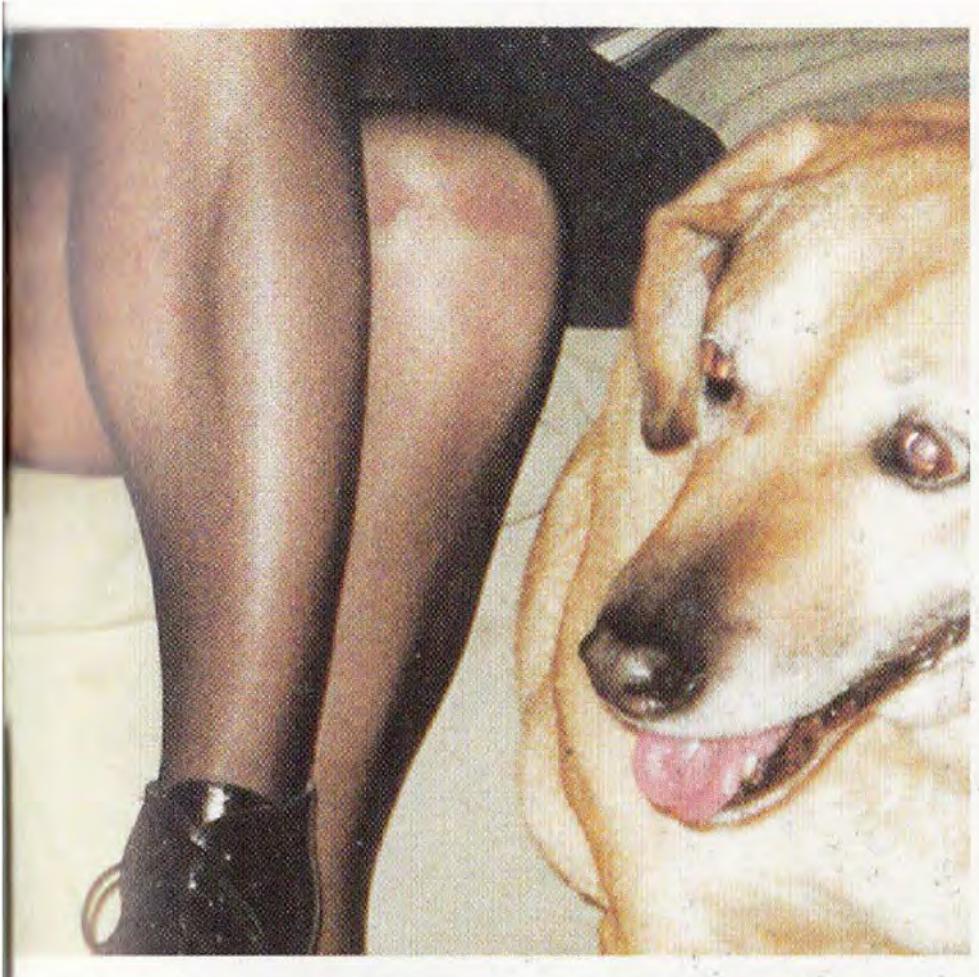
- “다른 것을 추구하면서도 무언가를 찾아내는 능력, 혹은 자신이 찾는 것보다 더 중요한 것을 발견했다는 사실을 인지할 수 있는 능력”
- Jean Jacques, *L'imprévu, ou, La science des objets trouvés.*

바술라르

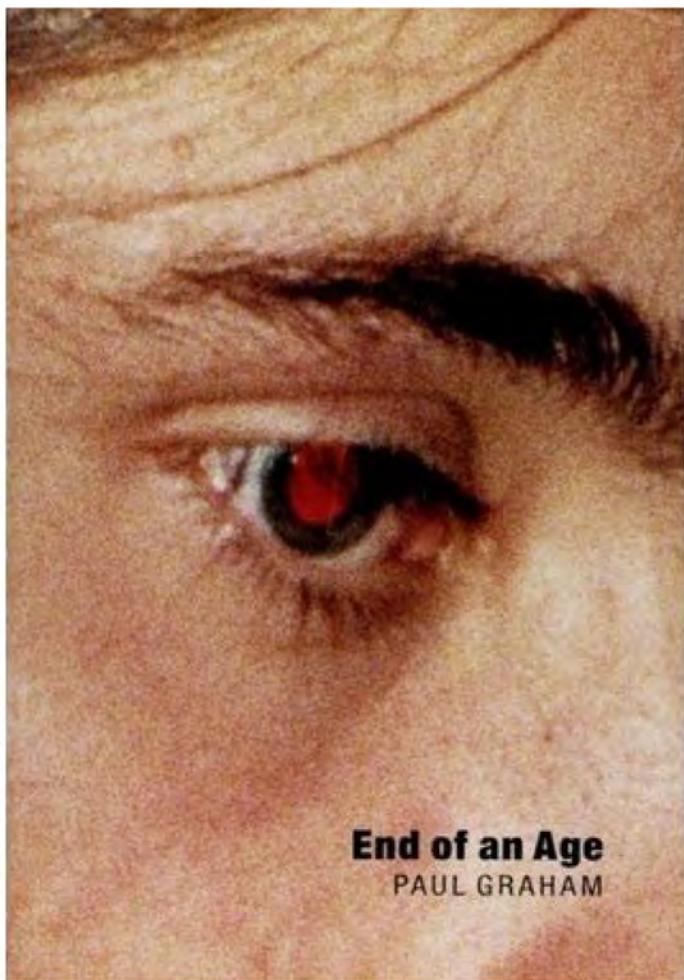
- “과학적 지식의 발전에 걸림돌로 인식된 소위 ‘장애물(obstacle)’은 오히려 지식의 형성과정 혹은 과학자의 인지과정에서 매우 중요한 지표이다.”



Philippe Durand,
« Fille, BM >>, de la série *Pharmacie*, 1999



Laurent Mulot,
« La belle et la bête »,
de la série *Sujet principal coupé : les deux yeux ne sont pas visibles*,
mars 1999



- “흐릿함, 잘못된 컬러, 사진기 흔들림, 적목현상, 잘못된 플래시와 같은 아마추어 사진 특징들을 이용한다.”





아마추어 실수기법

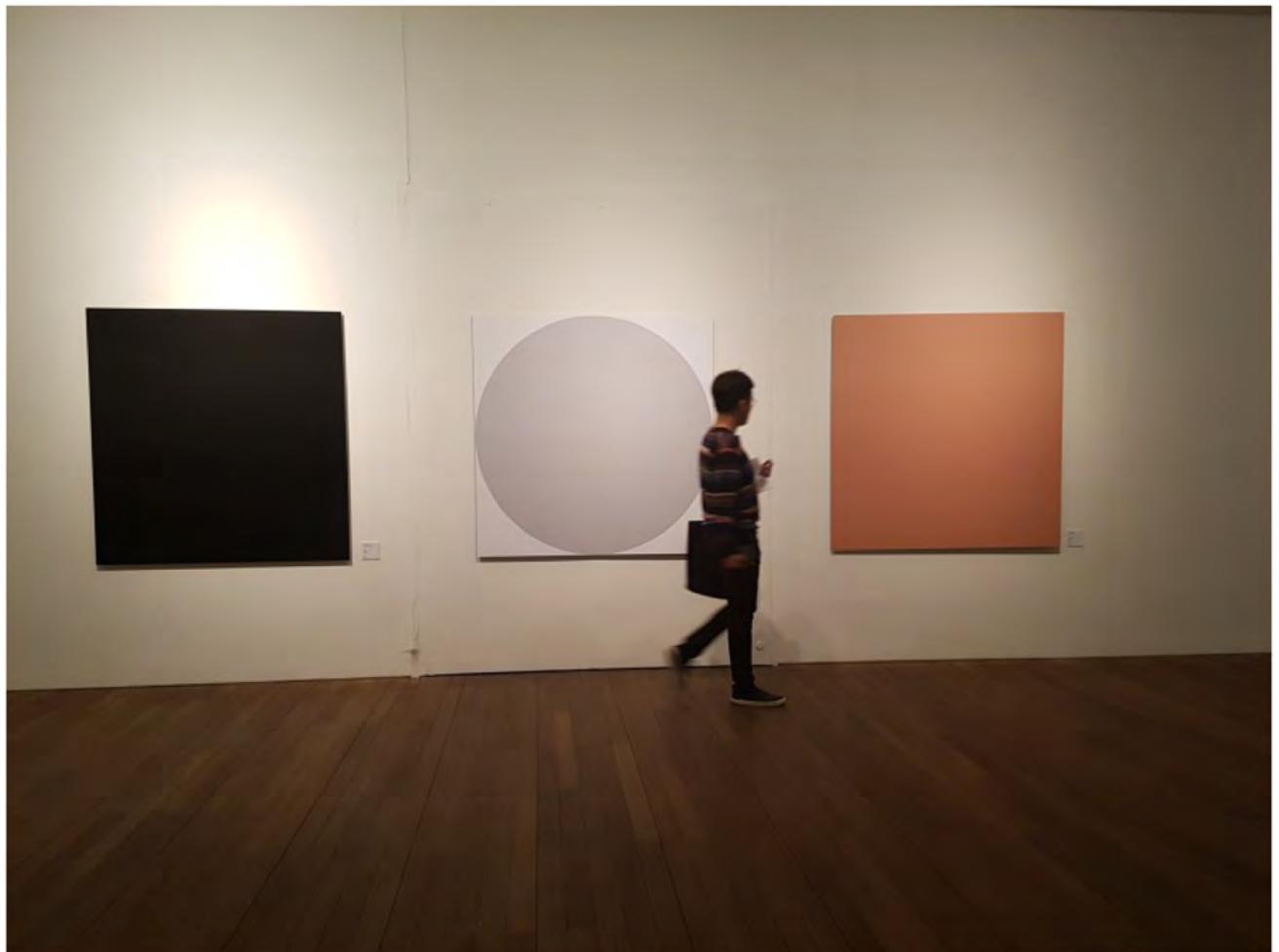




Bientôt chez lui, Mr. Dobard se plonge pendant 3/4 d'heure dans la plus profonde obscurité afin de faire un de ses beaux dessins que vous connaissez.

au bout de ses 3/4 d'heure, Mr. Dobard relève sa planche, mais l'opération se trouve totalement manquée à cause d'un tout petit malice au coin dans le ciel.

18he - 19th





Femina Sacer



카니발의 여신



2012 친구들 #2

얼굴의 철제 청대에는 큰 페가이 있으나 머리 색깔이 끼도하게 변형됐다.
생식 이미지를 만드는 고대상도 단계에서
스피일의 변화가 큰 폭으로 진행되었음을 알 수 있다.

머리에 정체불명의 얼룩이 생기나 기이한 형상이 만들어졌다.
이를 해결하기 위해 고안된 모양으로 얼룩이 생기나는 문제를
완전히 극복하지는 못한 것으로 보인다.

중심인물 주변에 위치한 다른 인물들의 모습이 흥미롭게 일그네져 있다.
기계가 주연 인물을 배경으로 인식하여 무자위로 합성한 결과다.



거울의 균열: 기계-의인화의 종말과 창의성

이재준
숙명여자대학교

거울의 균열: 기계-의인화의 종말과 창의성

Cracks in the Mirror:
The End of Machine-Anthropomorphism and Creativity

이 재 준(숙명여자대학교 인문학연구소 조교수)

2023년 6월 1일

제6회 인공지능인문학 국제학술대회
기술비평 심포지엄: 인공지능과 기계의 창의적 오류

[1]

기계의 창의성과 기계-의인화

1-1. 기계의 창의성

어떤 창의성?

creativity, the ability to make or otherwise bring into existence something new, whether a new solution to a problem, a new method or device, or a new artistic object or form. (Britannica)

창의성은 새로운 해결책, 새로운 방법 또는 장치, 또는 새로운 예술적 대상이나 형태를 만들거나 존재하게 할 수 있는 능력입니다.

창의성은 새로운 아이디어를 생각해내고, 새로운 해결책을 찾고, 새로운 것을 창조하는 능력입니다. 창의성은 모든 분야에서 중요합니다. 예술, 음악, 과학, 기술, 비즈니스 등 모든 분야에서 창의성이 필요합니다.

창의성은 타고난 능력이 아니라 배울 수 있는 능력입니다. 창의성을 키우기 위해서는 다양한 경험을 하고, 새로운 것을 시도하고, 실패를 두려워하지 말아야 합니다. 창의성은 우리 삶을 더 풍요롭게 만들 수 있는 능력입니다.

Plato thus set the stage for the approach to creativity that does not rely on trying to find an explanation that makes creations humanly predictable. In contrast to the idea of inspiration, Aristotle set forth an explanation of art as craft, according to which creating is a straightforward and productive process of imposing a preconceived plan on some material. Kant's contribution is found in his discussion of genius in The Critique of Judgment. Genius is the condition by which new rules are given to art. New rules are the new ways artworks are formed so that they initiate new styles. These rules or ways of forming works of art are prompted by ideas ("aesthetical ideas") that express or mean more than can be understood conceptually. Neither these aesthetic ideas nor the rules are concepts; they cannot be conceptualized.

Croce's and Collingwood's aesthetic theories stressed the crucial role of imagination (which must be distinguished from conceptual thought) as creative and integral to expressive activity in the creator's experiences.

The idea that creative activity can be traced neither to a source beyond human and natural activity nor to established regularities suggests that special attention should be given to the creative agent and the productive process. As a result, most studies in the twenty-first century have been focused on three distinct kinds of interests: (1) the person (the agent who works creatively), (2) the process itself, and (3) the product (the outcome of the process, which is what alerts us to something noteworthy about the character of the process that leads to a created product).

(Creativity, the Encyclopedia of Aesthetics, 2014, 2nd, ed. By Michael Kelly)



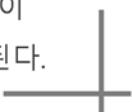
창의성은 새로운 것을 생각해내고, 새로운 해결책을 찾고, 새로운 것을 창조하는 능력이다. 창의성은 타고난 능력이 아니라 배울 수 있는 능력이다. 창의성을 키우기 위해서는 다양한 경험을 하고, 새로운 것을 시도하고, 실패를 두려워하지 말아야 한다. 창의성은 우리 삶을 더 풍요롭게 만들 수 있는 능력이다.

이 글에서 언급된 창의성의 개념은 다음과 같다.

- Plato: 창의성은 신성한 영감에서 비롯된다.
- Aristotle: 창의성은 예술가의 기술과 지식에서 비롯된다.
- Kant: 창의성은 새로운 규칙을 세우는 능력이다.
- Croce and Collingwood: 창의성은 상상력에서 비롯된다.

이러한 다양한 개념들은 창의성이 무엇인지에 대한 우리의 이해를 넓혀준다. 창의성은 단일한 개념이 아니라 다양한 요소들이 복합적으로 작용하는 결과물이다. 창의성은 신성한 영감, 기술, 지식, 상상력, 그리고 그 밖에도 많은 요소들에서 비롯될 수 있다.

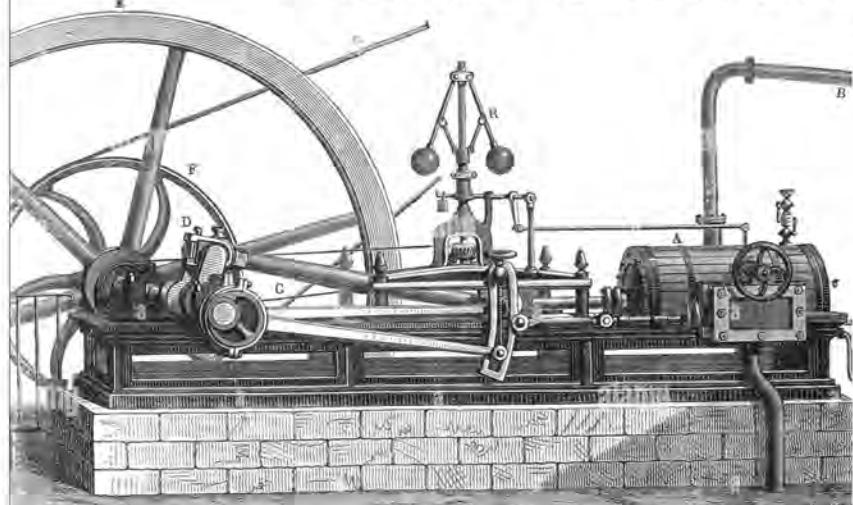
창의성은 우리 삶의 모든 영역에서 중요하다. 예술, 음악, 과학, 기술, 비즈니스 등 모든 분야에서 창의성이 필요하다. 창의성은 우리 삶을 더 풍요롭게 만들고, 문제를 해결하고, 새로운 것을 발견하는 데 도움이 된다.



기계가 창의적일 수 있나? 기계에 대한 일반적인 선입견

- 기계는 규칙을 물리적으로 반복하는 단단한 구축물이다.
- 기계는 설계에 따라 제작된 것이다.
- 기계는 목적에 순응한다. 즉 도구이다.

Levy Bryant's Machine-Oriented Ontology



'어떤 기계가 창의적일 수 있는가?'

- 지능적 기계가 그렇다.

Alan Turing, "Computing machinery and intelligence"(1950)
The **analogous question** as to whether a computer could think.

Thinkable Things / Copying Things

Weak AI : 인공지능은 인간이 생각하는 것처럼 작동한다.
Strong AI: 컴퓨터가 진정으로 생각할 수 있다.

Weak AI Cr : 인공지능은 인간이 창의적인 것처럼 작동한다.
Strong AI Cr : 컴퓨터가 진정으로 창의적일 수 있다.

- 인식 행위 능력=창의적 행위 능력



'어떤 기계가 창의적일 수 있는가?'

- 지능적인 것은 지적 능력이고, 안다는 것이다. 그런데 '체화된 인지'처럼 여러 종류가 있다.
- 인간의 행위가 창조적인 것으로 이해되려면, 적어도 다음과 같은 조건이 충족되어야 한다. 그 행위 혹은 결과가 놀랍고, 독창적이고, 자발적이어야 한다.
- 인공지능이 이런 창의성 조건 중 하나를 실현한다면, 어느 정도는 창의적이라고 볼 수 있다.

Paul, Elliot Samuel and Dustin Stokes, 2018, "Attributing Creativity", in Gaut and Kieran 2018: 193–210. Gaut, Berys and Matthew Kieran (eds.), 2018, Creativity and Philosophy, New York: Routledge. doi:10.4324/9781351199797



인공지능 기계의 창의성

Generative Art Systems(DALL-E, Midjourney, Stable Diffusion, VQGAN+CLIP etc.



Drew Harwell, "arts competition. Was it cheating?"(Washington Post, Sept. 2. 2022.)
<Théâtre D'opéra Spatial>(2022), Jason Allen's AI-generated art won first place in the Colorado State Fair arts competition.



Ai-Da with Portrait of the Robot(2021, Wiki)

- 미학적 투링테스트(인간주의적 기만을 통해 기존 예술 개념을 부정한다.)
- 심리적 정당화(창의성에 필요한 참신함을 포함한다.)
- 예술 제도론적인 정당화(창의성에 필요한 예술적인 특성 비교한다. Quasi-Artworks)



1-2. 기계-의인화

LLM의 '환각 현상'과 앵무새 효과

인공지능 언어모델은 자연언어를 실질적으로 이해하는 것이 아니다. 그것은 단순히 확률론을 기반으로 단어들을 조합한 결과값을 출력하기 때문에 그 기계는 단지 '확률적 앵무새 (Stochastic Parrots)' '일 뿐이다.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021, March). 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.



대형언어모델(LLM)은 '확률적 앵무새'이다. 오픈AI가 사용한 데이터가 무엇인지 알리지 않는 것은 위험한 일이다. 즉, 제어가 안 된다(Geoffrey Hinton). LLM의 위험을 완화하기 위해 웹에서 무차별적으로 데이터를 수집하는 대신 data sets을 골라내고 신중하게 문서화해야 한다.

Alan Turing, "Computing machinery and intelligence"(1950)
 The **analogous question** as to whether a computer could think.



Searle, John R. (1980), "Minds, Brains, and Programs", Behavioral and Brain Sciences 3: 417-457.

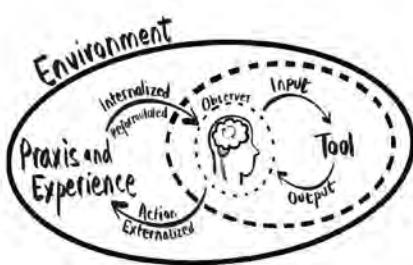


거울은 어디에

확률적 접근 너머, 입력된 것들로부터 창발한 어떤 세계 이해가 있다.

(Illya Sutskever, OpenAI)

자기귀환적 과정(autoregressive process)로부터 제2질서 사이버네틱스



Tilak, S., Glassman, M., Kuznetcova, I., & Pelfrey, G. L. (2022). Applications of cybernetics to psychological theory: Historical and conceptual explorations. *Theory & Psychology*, 32(2), 298-325.

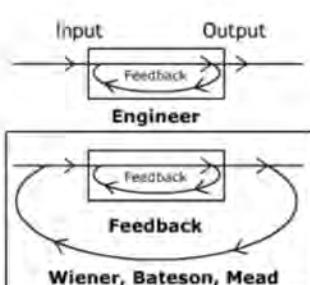
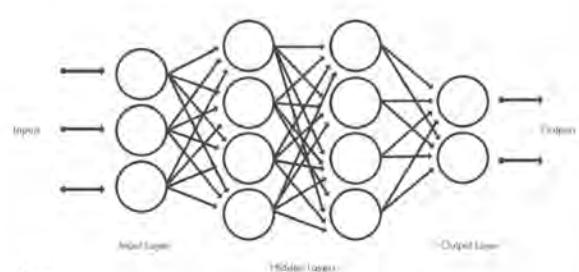


Diagram from Stewart Brand's 1976 conversation with Margaret Mead and Gregory Bateson, noting that they and Norbert Wiener understood themselves as participant observers in contrast to the detached "input-output" approach typical of engineering(Wikipedia)

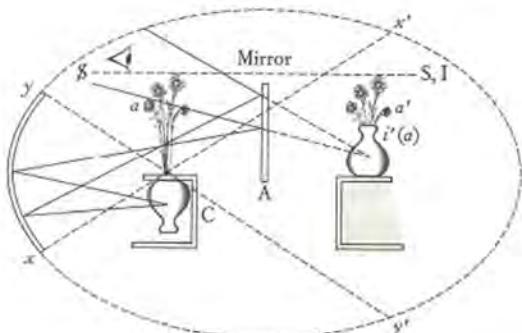


[2]

거울의 균열과 창의적 생산 다시 생각하기

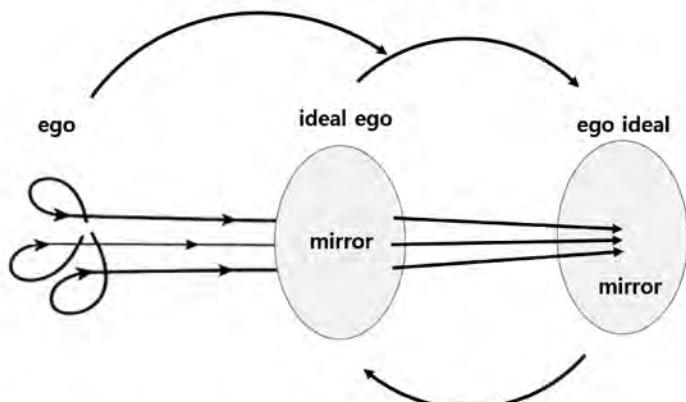
2-1. 거울의 균열

J. Lacan의 거울 시스템과 그것의 균열



Kittler, F. (1999). Gramophone, Film, Typewriter, trans. Winthrop-Young G and Wutz M. Stanford: Stanford University Press.

Process of Narcissism

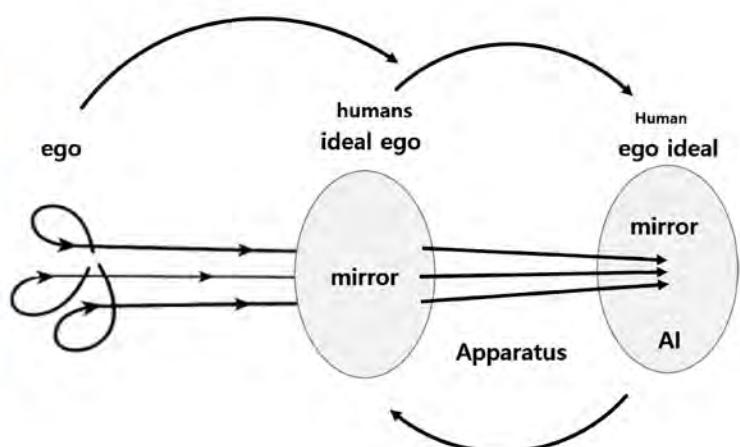


- Heimlich/ Unheimlich(Canny/ Uncanny)
- A Repetition of the Repetition Compulsion Itself. (A kind of meta-uncanniness)

기계-의인화의 종말, 거울 시스템의 균열



- 인공지능, 언어 모델, 자연언어의 불능,
상징들의 체계, 확률론적 모방

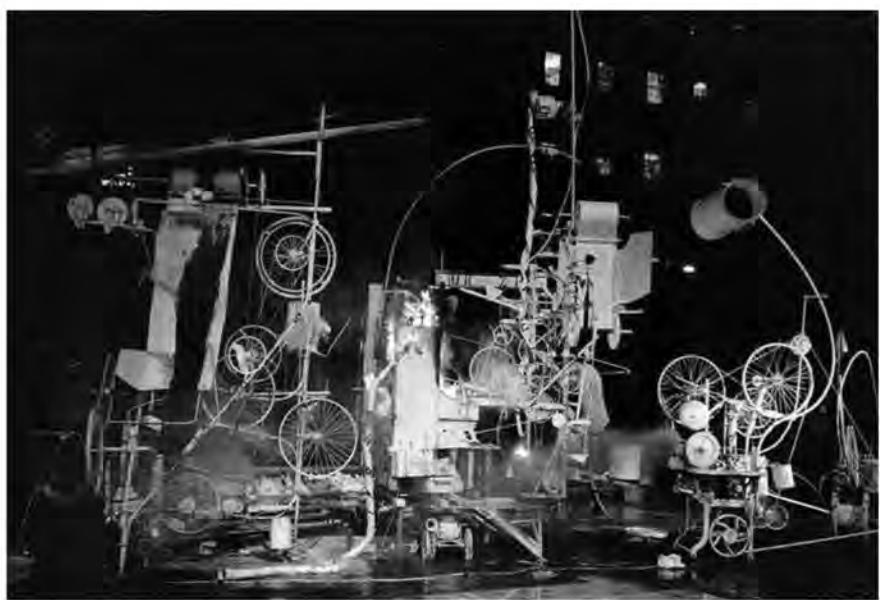


'Unheimlich' Recursiveness(Compulsive Repetition)

2-2. 창의적 생산을 다시 생각하기

Art as humanistic Production

Jean Tinguely, (Fragment from Homage to New York), 1960



Art as Post-humanist's humanistic Production



기계 예술 생산, 우발성의 효과, post-human uncanny
인간-AI의 중간 횡단 과정에서 폭로된 인간적인 것

Eunsu Kang, Sahn-Barni Poczos with Dall-E 2, 2022



감사합니다.