

LECTURE NOTES

Discourse, the Public Sphere & Disinformation

Habermas, information disorder and the transparency of
synthetic content

MODUL

**Introduction to the Ethics and Regulation of
AI**

SEMESTER

WS 26/27

VERSION

1.0

Prof. Dr. Markus Oermann

Instructor

markus.oermann@thws.de

Technische Hochschule Würzburg-Schweinfurt

Stand: July 23, 2026

Contents

1	Why democracy needs a public sphere	3
1.1	From the will of the people to public opinion	3
1.2	Speech acts and their validity claims	3
2	Discourse ethics: principles D and U	4
3	The digital public sphere and its discontents	5
3.1	From one-to-many to many-to-many	5
3.2	Filter bubbles and echo chambers: the empirical pushback	6
4	Information disorder: mis-, dis- and malinformation	7
4.1	Generative AI as an engine of information disorder	7
4.2	Deep fakes and synthetic reality	8
5	The regulatory response: transparency instead of prohibition	9
6	From discourse to transformation	9
7	References	10
7.1	Literature	10
7.2	Norms & Standards	10

The previous unit ended with the fear that predictive models let a few actors know and steer what many people think, and this unit asks the political question that follows: what happens to democratic opinion-formation when both the information and the interlocutors in public debate can be manufactured by machines? We begin with the classical theory of why democracy needs a functioning public sphere, move through the discourse ethics that tells us when a norm may count as valid, then confront the empirical reality of the digital public sphere, the vocabulary of information disorder, deep fakes and synthetic reality, and finally the regulatory answer that European law now offers.

01

KAPITEL

Why democracy needs a public sphere

1.1 From the will of the people to public opinion

Democracy means that political rule is tied back to the will of the governed. In Abraham Lincoln's Gettysburg formula it is "government of the people, by the people, and for the people", and in a parliamentary democracy elections are the mechanism that carries the majority will of the population into the legislative and executive branches. This presupposes something easy to overlook: that minority views can *become* majority views over time. That in turn requires the free formation of opinion, which above all means that opinion-formation is not controlled by the state or by those who hold power.

Jürgen Habermas gave this requirement its classical form in *Strukturwandel der Öffentlichkeit* (1962). In a modern mass society, he argued, the free formation of public opinion rests on a plurality of freely reporting journalistic and editorial media. These media perform a **public-watchdog** function: they are meant to supply a reliable base of information for forming opinions, and they serve at the same time as forums for the exchange of opinions and as an infrastructure that makes deliberation possible at all. The public sphere, on this account, is not a luxury of democracy but one of its enabling conditions.

DEFINITION

The public sphere (Habermas 1962)

The communicative space, carried by free and plural media, in which private citizens form and exchange public opinion free of state control; its watchdog and forum functions supply the information base that democratic opinion-formation presupposes.

1.2 Speech acts and their validity claims

Before we can ask when public discourse *succeeds*, we need a theory of what we do when we speak. Linguistics has shown that utterances always carry an element of action: we do things with

words, we communicate in and through **speech acts**. As a communication community we attach different **validity claims** to different kinds of speech act. To every communication we attach a claim to *comprehensibility*; to expressive utterances (representativa) a claim to *truthfulness*, meaning inner sincerity; to normative utterances (regulativa) a claim to *normative correctness*; and to factual assertions (constativa) a claim to *truth*. Truth, on this view, is not a property of statements in the abstract but a validity claim we raise for a particular class of statements and can be asked to redeem.

EXERCISE · TERMS

An expressive utterance raises a claim to truthfulness, a normative utterance a claim to correctness, and a factual assertion a claim to truth, while every utterance at minimum raises a claim to comprehensibility.

02

KAPITEL

Discourse ethics: principles D and U

These validity claims are redeemed, if at all, in **discourse**: the rational exchange of arguments. Habermas and Karl-Otto Apel model this against the counterfactual ideal of an *ideal communication community*, in which the only force is the force of the better argument. Discourse theory distinguishes two families. In **theoretical discourses** we argue about constativa, about what is *true*, as in science. In **practical discourses** we argue about regulativa, about what is *normatively correct*, as in politics. This distinction matters for AI ethics, because it locates the disputes of this course, over fairness, over legitimate regulation, squarely in the practical register, where the test is not correspondence to facts but the assent of those affected.

From this Habermas and Apel derive **discourse ethics**, which offers two conditions for the validity of a norm. The **discourse principle (D)** holds that valid norms are precisely those to which all those affected could agree as participants in a rational discourse. The **universalisation principle (U)** adds a test that any single participant can apply in thought: a norm is valid only if the consequences and side effects that would follow from its general observance can be accepted by everyone affected. Principle U is the discourse-ethical successor to Kant's categorical imperative, but it relocates the test from the solitary conscience to an actual or virtual dialogue.

DEFINITION**Principles D and U (Habermas, Apel)**

Principle **D**: valid norms are those all affected parties could accept as participants in a rational discourse. Principle **U**: a norm is valid only if the consequences of its general observance are acceptable to everyone affected.

QUICK CHECK

According to the universalisation principle (U), a norm is valid when:

it maximises the aggregate welfare of society

it is enacted by a democratically elected legislature

the consequences of its general observance can be accepted by all those affected

it follows logically from a higher moral law

The reason discourse ethics belongs in a unit on the *public sphere* is now visible. Principles D and U are only meaningful if a real discourse is at least orientated to the the ideal one: if the affected parties can actually speak, be heard and revise their views. That is exactly what a functioning public sphere is supposed to guarantee, and exactly what the digital transformation of that sphere puts in question.

03

KAPITEL

The digital public sphere and its discontents

3.1 From one-to-many to many-to-many

Mass media (print, radio, television) organise public communication in a **one-to-many** structure: an editorial office decides which information is published and can therefore be perceived at all. Social media invert this into a **many-to-many** network structure in which, at least in principle, everyone can address everyone. The gatekeeping does not disappear, however; it changes hands. Which posts are seen by whom is now determined chiefly by search and, above all, **feed algorithms** that select, sort and rank. As media use shifts to social platforms and these become the first point of access to information, control over what is perceived migrates from the editorial offices of media companies to the platform operators, whose selection logic is optimised for engagement rather than for the informational needs of a democratic public. Habermas's diagnosis, first published in 1962, has only sharpened in this platform age.

3.2 Filter bubbles and echo chambers: the empirical pushback

Two related worries dominate the debate. The **filter-bubble** thesis holds that algorithmic personalisation increasingly shows individual users content that matches and reinforces their existing views, sealing them into a private informational world. The **echo-chamber** thesis, associated with Eli Pariser, adds a social mechanism: people on social media tend to talk mainly with the like-minded, forming groups closed off from one another. Both would corrode the shared information base that principles D and U require.

The pedagogically important point is that the strong versions of these theses are not well supported by the evidence. Personalisation-driven filter bubbles remain, on the best current reading, empirically unproven as a dominant effect (see Guess et al. and the wider literature). Echo chambers, similarly, look less like a pure product of social-media architecture and more like a reflection of pre-existing groups with shared worldviews. Gounaridis and Newell (2024), studying the geography of climate-change denial in the United States, illustrate this: denial clusters spatially, but the clustering tracks partisan and community structure rather than being manufactured wholesale by feeds. The lesson is one of intellectual discipline. The digital public sphere is genuinely fragmented and polarised in ways that matter for democracy, but the causal story is not the tidy technological determinism that “filter bubble” suggests, and a serious ethics of the public sphere must resist the temptation to over-attribute.

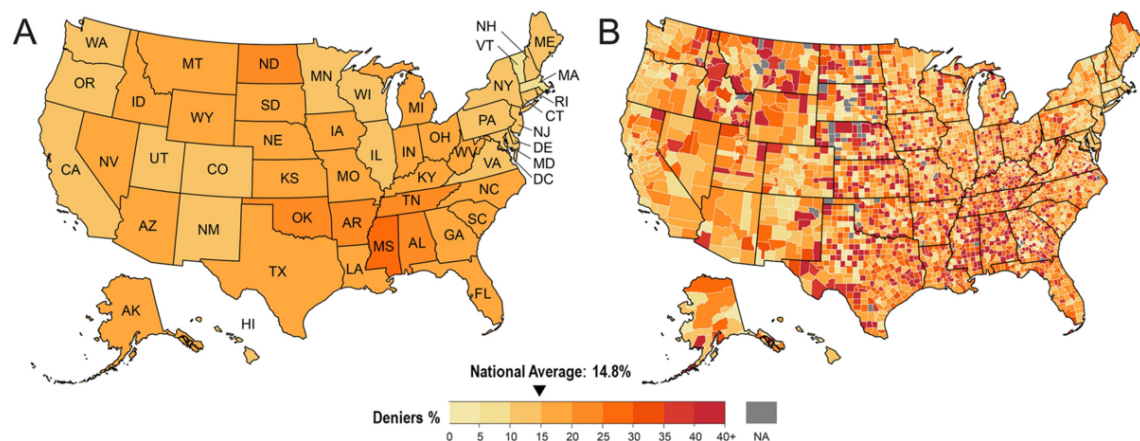


Figure 1: Climate-change denial by U.S. county: panel A shows the share of “deniers” against a national average of 14.8%; panel B maps the strong spatial clustering. The pattern tracks partisan and community structure rather than a pure social-media effect (source: Gounaridis, D. & Newell, J.P. (2024). The social anatomy of climate change denial in the United States. *Scientific Reports* 14, 2097, Fig. 1, <https://doi.org/10.1038/s41598-023-50591-6>, CC BY 4.0).

DEEP DIVE Deep dive: why the correction matters ethically

It would be convenient for regulators if algorithms simply built the walls of the echo chamber, because then removing the algorithm would remove the harm. The empirical picture is less obliging. If fragmentation reflects offline group structure amplified, but not created, by platform dynamics, then a purely technical fix addresses only part of the problem, and interventions that assume otherwise may disappoint. Thiel’s structural analysis (2018) makes the deeper point: even without full informational isolation, it is enough that platform dynamics

foster polarisation and erode trust for democratic harm to follow. Honesty about causation is therefore not academic hair-splitting; it is what keeps regulation proportionate and effective.

04

KAPITEL

Information disorder: mis-, dis- and mal-information

Beyond homogeneous, mutually closed sub-networks, the second great challenge to successful public discourse is **false or harmful information**. The field has converged on a three-way distinction that turns on two questions: is the content *false*, and is it spread with an *intent to harm*? **Misinformation** is false information, untrue statements of fact, spread without the intent to harm, typically in honest error. **Disinformation** is false information spread deliberately and knowingly to deceive or damage. **Malinformation** is a residual category of otherwise harmful content, hate speech, terror propaganda, abuse material, or genuinely true information weaponised to injure, where the damage does not hinge on any statement being false. All three can occur in any medium: text, photo, audio, video, drawing. Sort the scenarios in the classifier below along the two axes and watch how the same fact pattern changes category as intent changes.

DEFINITION

Mis- / dis- / malinformation

Misinformation: false content, no intent to harm (honest error). *Disinformation*: false content spread deliberately to harm. *Malinformation*: harmful content, including true information weaponised to injure, where harm does not depend on falsity.

4.1 Generative AI as an engine of information disorder

Generative AI touches all three categories at once. By its very construction it produces **misinformation**: hallucination is not a bug at the margin but a consequence of the stochastic mechanism, so that a system can generate untrue statements of fact in perfect good faith and pass them to a user who does the same. It can be steered to create explicit **malinformation**, hate messages, depictions of violence, and abuse material, which is why keeping such output out of a deployed system depends on a large human workforce of data trainers and moderators, a labour question the next unit takes up directly. And it collapses the cost of **disinformation** by making persuasive falsehoods cheap, scalable, customisable and, increasingly, interactive.

4.2 Deep fakes and synthetic reality

The manipulation of media is almost as old as recording itself, but generative AI marks a threshold. It permits the creation of **deep fakes**: deceptively real yet wholly inauthentic images, audio and video of people in situations that never occurred. Prominently discussed cases include a fabricated image of one head of state kneeling before another, a staged “arrest” of Donald Trump, a fake image of a burning Pentagon that briefly moved stock markets, and non-consensual sexual imagery of Taylor Swift. Emilio Ferrara (2026) argues that the right unit of analysis is no longer the individual fake but **synthetic reality**: coherent, interactive and personalisable information environments in which content, identities and social interactions are jointly fabricated and mutually reinforcing. The genuinely new capacity is not the forged artefact but the ability to fabricate an entire *context of belief*, who is present, what evidence exists, which claims circulate and how they are socially validated.

Ferrara’s sharpest thesis is the **generative AI paradox**: the more ubiquitous synthetic content becomes, the more *rational* it becomes for a society to discount digital evidence in general, which raises the cost of truth and hands strategic actors a “liar’s dividend”, since genuine evidence can now be dismissed as fake. This asymmetry also explains why *debunking* is a weak remedy. There is a time gap in which serious damage occurs before a fake is corrected, as the Pentagon case showed; corrections do not reliably reach everyone who saw the original; misinformation is especially sticky when it confirms the recipient’s worldview (confirmation bias); and, psychologically, a correction seen later is simply less effective, because something of the false first impression tends to remain.

CASE STUDY

Case: the market-moving fake

A photorealistic but fabricated image of a large explosion at a government building spreads on social media during trading hours. Before news organisations can verify that it is a deep fake, automated and human traders react and an index dips sharply. By the time the debunking circulates, some damage is done. Which categories apply, and why is the after-the-fact correction only a partial remedy?

Solution. The image is a fabricated statement of fact, published to deceive, so it is **disinformation** delivered as a deep fake; if the aim were purely to injure a named party with harmful but non-false material, the label would shift toward **malinformation**. The correction fails to fully repair the harm for the four reasons the debunking literature identifies. The **time gap** means the financial damage is realised before verification arrives. The **outreach** problem means the correction does not reach every viewer of the original. **Confirmation bias** means recipients already primed to expect instability absorb the fake more readily than the correction. And the **primacy** effect means a later correction is psychologically weaker than the first impression. This is exactly the dynamic Ferrara’s generative-AI paradox predicts: as fakes proliferate, the epistemic cost of establishing what is real rises for everyone, which is why the regulatory response focuses on *provenance* and *labelling* upstream rather than on debunking downstream.

05

KAPITEL

The regulatory response: transparency instead of prohibition

Faced with synthetic content, providers of large generative systems have reacted with ad hoc technical restrictions, especially before elections, and platforms such as YouTube and Meta have introduced labelling requirements. The European legislator's structural answer is not to ban synthetic content, which would collide with communication freedoms, but to make it **recognisable**: this is the logic of Art. 50 AI Act, the limited-risk transparency duty examined in detail in the unit on the EU AI Act. The content itself is not governed, only its transparency, a real step toward closing what Ferrara calls the **provenance gap**, the absence of a reliable chain of custody for digital media, though it remains structurally limited: watermarks can be fragile under compression, re-encoding and cross-model remixing, and the "analog hole" (re-recording a screen) defeats embedded signals. Transparency, in short, mitigates the harm of synthetic reality without dissolving it, which is why the deeper answers still point back to platform governance, institutional workflow redesign and public epistemic resilience.

06

KAPITEL

From discourse to transformation

We have traced a single argument across this unit: democracy depends on a public sphere in which discourse can approximate the ideal that principles D and U presuppose, the digital transformation of that sphere strains this condition through fragmentation and, more acutely, through information disorder and synthetic reality, and European law responds not by banning content but by trying to keep its origins transparent. That regulatory answer, like the moderation of harmful output, rests on a great deal of hidden human labour and consumes real physical resources, and it is to the transformation of work and the sustainability of AI that the next unit turns.

07

KAPITEL

References

7.1 Literature

- Ferrara, E. (2026): *The Generative AI Paradox: GenAI and the Erosion of Trust, the Corrosion of Information Verification, and the Demise of Truth*. arXiv:2601.00306.
- Gounaridis, D. & Newell, J. P. (2024): The social anatomy of climate change denial in the United States. *Scientific Reports* 14 (1), 2097. <https://doi.org/10.1038/s41598-023-50591-6>
- Guess, A. M., Nyhan, B., Lyons, B. & Reifler, J.: on the empirical limits of the filter-bubble and echo-chamber theses; see the summary of the literature in the works cited here.
- Habermas, J. (1962): *Strukturwandel der Öffentlichkeit: Untersuchungen zu einer Kategorie der bürgerlichen Gesellschaft*. Luchterhand, Neuwied.
- Habermas, J. (1991): *Erläuterungen zur Diskursethik*. Suhrkamp, Frankfurt am Main.
- Lorenz-Spreen, P., Lewandowsky, S., Sunstein, C. R. & Hertwig, R. (2020): How behavioural sciences can promote truth, autonomy and democratic discourse online. *Nature Human Behaviour* 4, pp. 1102-1109. <https://doi.org/10.1038/s41562-020-0889-7>
- Pariser, E. (2011): *The Filter Bubble: What the Internet Is Hiding from You*. Penguin Press, New York.
- Thiel, T. (2018): Digitalisierung: Gefahr für die Demokratie? *Politikum* 4 (4), pp. 4-11.

7.2 Norms & Standards

- Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act), in particular Art. 50 (transparency obligations for certain AI systems). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=OJ:L_202401689