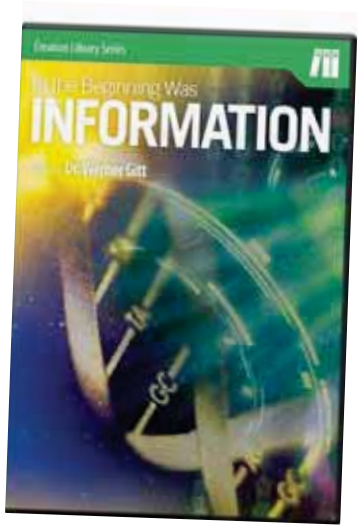
# Informasjon

*Av Professor John C. Lennox, Universitet i Oxford*

## saken kort

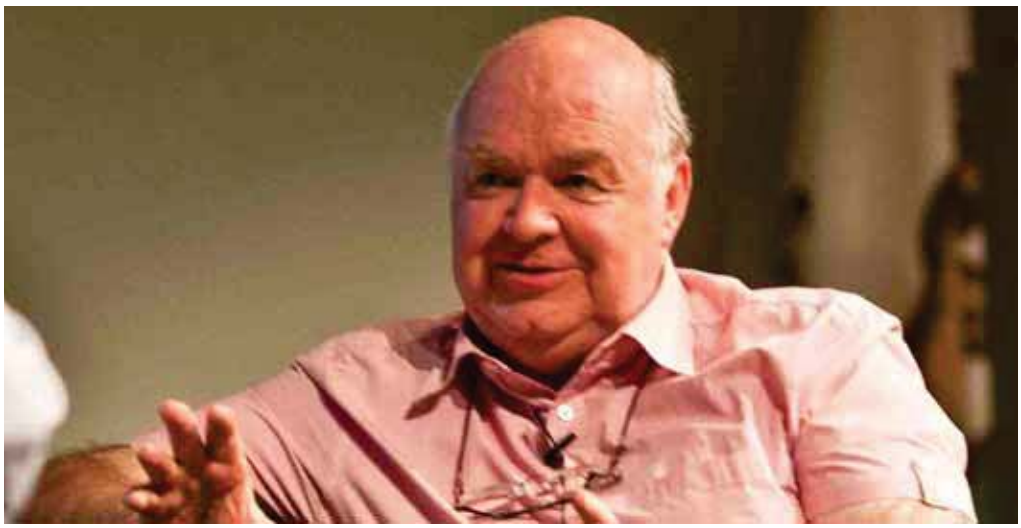
Vi gjengir her et utdrag av kapitlet om Informasjon fra boken *God's undertaker. Has science buried God?* Origo regner med å utgi boken i løpet av året. Teksten i denne artikkelen er oversatt av Jon Kvalbein, og baserer seg på 1. utgave av boken.

### Hva er informasjon?




I dagligtale bruker vi ordet "informasjon" for å beskrive noe som vi nå vet som vi ikke visste fra før. Vi sier at vi har mottatt informasjon. Det finnes mange metoder til å formidle informasjon: gjennom alminnelig skrift, gjennom tegnspråk, ved kryptiske koder osv. Problemet dukker opp dersom vi forsøker å kvantifisere informasjon. Men informasjonsteorien har gjort store fremskritt. Dette er av stor betydning når vi nå skal betrakte det vi har kalt genetisk informasjon.

La oss starte med å se på den intuitive forestilling vi har om at informasjon minker vår usikkerhet. Vi kan tenke oss at vi kommer til et lite hotell der vi har reservert plass. Vi oppdager at det bare finnes åtte rom. Dersom vi antar at alle rom er like og at vi ikke har reservert et spesielt rom, er sannsynligheten 1 til 8 for at vi har fått et bestemt rom. Sannsynligheten er et klart mål på vår usikkerhet. Dersom vi får beskjed om at vi er plassert på Rom 3, forsvinner usikkerheten. En fremgangsmåte for å måle den informasjonen vi har fått, kunne være å finne ut det minste antall "ja og nei"-spørsmål som ville være nødvendig for å finne ut hvor vi skulle bo. Tenker vi oss litt om, ville vi finne at svaret er 3. Vi sier at vi har mottatt tre "bits" av informasjon, eller at vi trenger tre bits informasjon for å kunne spesifisere vårt rom. Vi merker oss at 3 er den potensen vi må opphøye 2 i for å få 8 (det vil si $2^3 = 8$). Eller for å snu litt på det, 3 er logaritmen til 8 med 2 som grunntall (det vil si $3 = \log_2 8$). Det er lett å generalisere denne argumentasjonen og se at dersom det var n rom på hotellet, ville den informasjonsmengden som trengs for å spesifisere ett bestemt rom være $\log_2 n$.



Tenk nå på et tekstbudskap som er skrevet på engelsk, som vi skal betrakte som et språk skrevet i setninger som består av ord og mellomrom, slik at vårt "alfabet" består av 26 bokstaver pluss "mellomrom". I alt 27 symboler er da nødvendig. Hvis vi venter på et budskap på vår mobiltelefon, kan vi i prinsippet tenke oss at sannsynligheten for å motta et hvilket som helst symbol er 1/27. Informasjonen som blir addert ved hvert tekstsymbol er da $\log_2 27$ (tilnærmet 4,76). Informasjonen som overføres ved en tekst som er m symboler lang, er da $m \bullet \log_2 27$ (tilnærmet $4{,}76 \bullet m$).

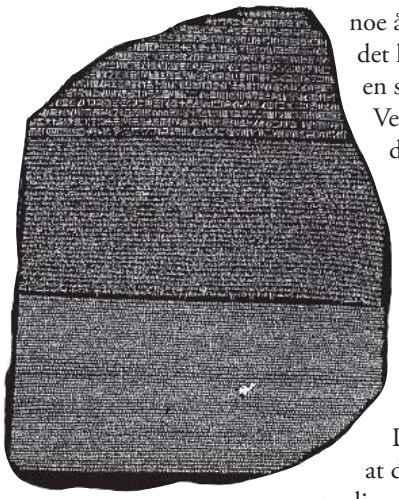
Vi merker oss her at mengden av informasjon som overføres er relativ til den kjente størrelsen av "alfabetet". Vi vet for eksempel at en tekstmelding kan inneholde tall i tillegg til bokstaver og "mellomrom". Da vil "alfabetet" ha 37 symboler. Og da vil informasjonsmengden knytte til hvert symbol være $\log_2 37$ (tilnærmet 5,2).

Overalt her spiller tallet 2 en viktig rolle. Faktum er at det symbol-"alfabetet" som brukes i datamaskiner, bare består av de to symbolene 0 og 1. Det er lett å innse at 2 er det minste antall symboler som trengs for å innkode et hvilket som helst alfabet. For eksempel: Dersom vi tenker på det engelske alfabet med 26 bokstaver pluss mellomrom, kan en streng på 5 symboler være nok til å innkode alle disse ($2^5 = 32 > 27$). Da vil vi kunne sette A = 00001, B = 00010, C =00011 osv.

### Syntaktisk og semantisk informasjon

Vi vil nå innføre en meget viktig tanke som det noen ganger kan være vanskelig å få tak i med en gang. Anta at vi får følgende budskap på vår mobiltelefon: ZXXTRQ NJOPW TRP. Dette budskapet består av 16 symboler. Etter vanlig kalkyle skulle dette føre til et informasjonsinnhold på $16 \bullet \log_2 27$ bits. Til dette kan du si: "Stopp en halv, dette er absurd, for jeg har ikke mottatt noe budskap i det hele tatt. Det finnes ingen

informasjon i dette vrøvlet." Nå kan naturligvis budskapet foreligge i kodet form. Kan hende dreier det seg om et hemmelig budskap. La oss anta at dette ikke er tilfelle. Hva da? Vi må da erkjenne at "informasjon" i den betydning vi har diskutert så langt, ikke har noe å gjøre med "mening" i det hele tatt. Vi kaller dette en syntaktisk informasjon. Ved første øyekast synes dette å være i strid med vår intuisjon sett i lys av våre daglige erfaringer. Derfor er det nødvendig å betrakte dette nærmere. Anta at du blir fortalt at det kommer en melding på din mobiltelefon. Du blir også fortalt at du kan motta fire mulige symboler (^*# ¤) og at meldingen består av fem symboler. Du titter på skjermen, og leser ^##¤*. Hvor mye "informasjon" har du mottatt? Vel, ingen – i den forstand at du har peiling på hva dette betyr. Ja, du kan ikke vite om dette betyr noe i det hele tatt. Men i syntaktisk forstand har du mottatt informasjon. Det er fire mulige symboler. Så sannsynligheten for at du kan få en av dem er ¼. Og informasjonen som gis ved hvert symbol er 2 bits. Hele budskapet består av 5 symboler som innholder 10 bits. Sagt på en annen måte: Hvis vi teller opp hvor mange mulige "budskap" (strenger på fem symboler) du kan formidle, blir dette $2^{10}$. Du vet nå hva budskapet er (ikke hva meningen er). Du visste ikke dette på forhånd. I den betydningen, har du mottatt informasjon.



Tenk igjen på hverdagens elektroniske kommunikasjon gjennom en kabel, for eksempel en vanlig telefonlinje. Til enhver tid vil ulike former for "informasjon" være på vei gjennom den. Det kan være vanlige stemmer, fax-meldinger, datakommunikasjon – alle slags strømmer av elektroniske "symboler". Noen vil ha mening for noen mennesker, men ikke for andre (en person som snakker kinesisk vil ikke gi noen semantisk mening for en som ikke snakker kinesisk). Andre strenger av tilfeldige symboler er uttrykk for støy på linja. De er fremkalt av tilfeldige elektroniske årsaker og har ingen mening i det hele tatt.

En kommunikasjonsingeniør er ikke interessert i meningen i det som formidles via kanalen. Hun er ikke opptatt av de enkelte strengene av symboler som blir transportert, men av helt andre forhold som: Hvor stor er kabelens kapasitet? Hvor mange symboler kan sendes gjennom den i løpet av et sekund? Hvor stor er påliteligheten – hva er for eksempel sannsynligheten for feil i symbolene på grunn av støy på linja? Hva er muligheten for å rette opp slike feil? Alt dette opptar oss alle. Mange er frustrert over at datakommunikasjonen tar lang tid i hjem som ikke har installert bredbånd.

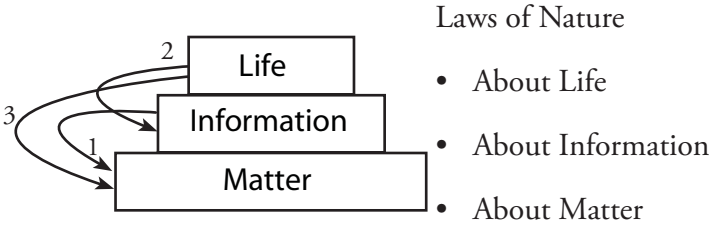
Måling av syntaktisk informasjon er derfor meget viktig. Teorien som er knyttet til dette er kalt Shannons teori om informasjon, etter Claude Shannon som utviklet den og beviste noen matematiske setninger om kapasiteten til en støyfull linje. Her er grunnlaget for den teori om kommunikasjon som samfunnet avhenger av i dag.

La oss ta for oss et annet eksempel fra hverdagen for å være sikre på at vi har fått tak i hva det dreier seg om. Vi går inn på et bibliotek og spør etter en bok om nephrologi. Bibliotekassistenten har aldri hørt om nephrologi. Men som en streng av symboler, innholder ordet "nephrologi" $10 \bullet \log_2 7$ bits av informasjon. Og hvis du gir assistenten denne informasjonen, kan hun skriver ordet inn i datamaskinens indekssystem og få som tilbakemelding at du kan lete i biblioteket under Medisin 46, der du vil finne 3 bøker om emnet. Dette innebærer at hun fungerer som en "kanal" for å formidle informasjon til indekssystemet, selv om symbolstrengen "nephrologi" ikke har noen som helst semantisk mening for henne.[1]

Å måle semantisk informasjon er et mye vanskeligere problem å behandle matematisk. Ingen god metode er funnet hittil. Dette bør neppe være overraskende. For det har å gjøre med at en tekst er meget avhengig av sin sammenheng. Hvis du ser meg motta en melding på mobiltelefonen med budskapet "JA", kan du fort tenke deg at dette er svaret på et spørsmål jeg har stilt. Men du kan ikke vite om dette spørsmålet er "Har du billett til fotballkampen i kveld" eller "Vil du gifte deg med meg?" Meningen av en tekst kan ikke bestemmes uten forhåndskjennskap til tekstens sammenheng. Med andre ord: Det trengs mye mer informasjon for å kunne tolke en gitt informasjon.

### DNA og informasjon

La oss nå anvende noe av denne tenkningen på molekylærbiologien. Tenk på den streng av "bokstaver" som vi finner i det kjemiske alfabetet til DNA-molekylet. Anta at du er en molekylærbiolog og vet noe om hva strengen av bokstaver "betyr" i den forstand at du kan dele den opp i gener og si hvilke proteiner de er kodet for osv. Det betyr at for deg har strengen en semantisk dimensjon. For deg uttrykker DNA presis samme slags spesifisert kompleksitet som et språk, fordi rekkefølgen på bokstavene i et gen spesifiserer rekkefølgen av aminosyrer i proteinene.




Men dette er ikke tilfelle for meg. Jeg ser strengen bare som en lang liste med meningsløse symboler ACGGTCAGGTTCTA … Likevel er det helt greit for meg å snakke om at jeg kjenner informasjonsinnholdet i symbolstrengen i syntaktisk betydning eller etter Shennons teori. Ja, til tross for at jeg ikke forstår "meningen" i strengen, kan jeg finne ut eksakt hvor mye syntaktisk informasjon du må gi meg for at jeg skal kunne reprodusere strengen nøyaktig. Det genetiske alfabetet består av fire bokstaver, slik at hver bokstav kan skrives ut

(eller sendes til min datamaskin) som to bits informasjon. For eksempel vil DNA i et menneskelig genom som er omtrent 3,5 milliarder bokstaver lang, inneholde omkring 7 milliarder bits av informasjon. Hvis jeg får dette, kan jeg skriver ut DNA uten å ha noen som helst innsikt i "meningen" i det jeg har skrevet ut.

Det er et viktig aspekt ved genom-forskningen å finne spesifiserte mønstre som blir gjentatt i et gitt genom eller å finne spesifiserte sekvenser som er felles for flere genomer. Årsaken til at man ser etter en spesifisert sekvens kan være av semantisk karakter. Men den reelle dataundersøkelsen i de store databaser som genomene representerer, er beskjeftiget med syntaktisk informasjon.

## Kompleksitet

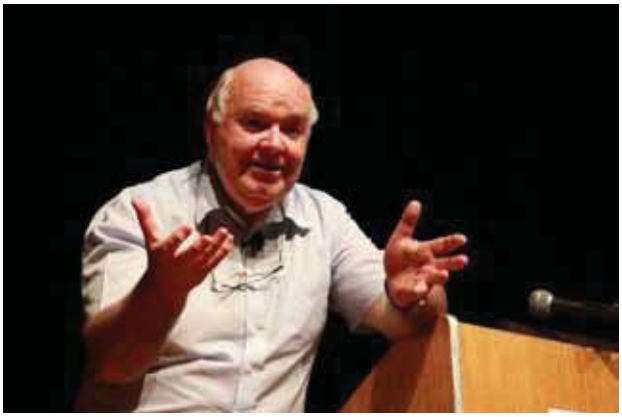
Så langt i dette kapitlet har vi ikke nevnt begrepet kompleksitet. Men vi kan umiddelbart forstå at det faktum at det menneskelige genomet inneholder 7 milliarder bits, gir oss litt idé om dets kompleksitet. Men bare litt. Tenk for eksempel på følgende binære streng: 001001001001001001001001… La oss anta at den fortsetter til vi har i alt 6 milliarder tegn (vi ønsker er tall som er delelig med 3). Da kan vi se, ut fra vårt perspektiv så langt, at strengen inneholder 6 milliarder bits. Er den da nesten så kompleks som det menneskelige genom? Langt ifra. For vi ser straks at strengen består av et repetert mønster – trippelet 001 blir gjentatt gang etter gang. På en måte kan vi si at all informasjonen i strengen kan formes i utsagnet: "Gjenta trippelet 001 to milliarder ganger." Denne mekaniske repetisjonen er et eksempel på hva matematikerne kaller en algoritme – den type prosesser som et dataprogram er konstruert for å utføre. I dette tilfellet kunne vi for eksempel skrive et enkelt program på følgende måte: "For n = 1 til 2 milliarder, skriv 001. Stop." For å skrive dette programmet, trenger jeg bare 44 tastetrykk. Og det blir straks klart at dersom vi tenker på 44 som "lengden" av programmet, gir dette oss et mye mer nøyaktig inntrykk av informasjonsmengden i strengen av binære tall enn det vi får av dens virkelige lengde på 6 milliarder tegn.

Et annet eksempel som lett vil får fram poenget er følgende: Betrakt strengen av bokstaver ILOVEYOUILOVEYOUILOVEYOUILOVEYOU … og anta at strengen består av 2 milliarder gjentagelser av de tre ordene I LOVE YOU. Det er tydelig at informasjonen (i semantisk betydning denne gangen) i strengen er gitt i de tre første ordene (selv om noen vil hevde at gjentagelsene understreker budskapet). I alle fall kunne vi skrevet den fulle semantiske informasjonen ved dataprogrammet: "For n=1 til 2 milliarder, skriv ILOVEYOU. Stop." Vi kunne derfor fått et bedre mål på informasjonsinnholdet ganske enkelt ved å telle antall bits av syntaktisk informasjon som finnes i dette korte programmet fremfor å telle antall bits i den lange teksten.

## Algoritmisk informasjonsteori

Denne "sammenpressingen" av en gitt symbolstreng (binære tegn, bokstaver, ord osv) slik at den opptar mindre plass, som utføres av et dataprogram, er en grunnleggende tanke bak det som kalles algoritmisk informasjonsteori. Ordet "algoritme" stammer fra navnet til matematikeren Mohammed Ibn-Musa Al-Khwarizmi som arbeidet i det berømte visdomshuset i Bagdad i det niende århundre. En algoritme er en effektiv prosedyre for å få noe trinnvis utført med et endelig antall trinn. Formelen $x = (-b \pm \sqrt{(b^2 – 4ac)})/2a$ gir oss en effektiv prosedyre for å beregne røttene til annengradslikningen $ax^2 + bx + c = 0$, der a, b, og c er tall. Dette er derfor en algoritme. På liknende måte er dataprogrammer (software) logaritmer som gjør datamaskinen (hardware) i stand til å utføre sin informasjonsbehandling. I alminnelighet vil dataprogrammer ta i bruk mange algoritmer, der hver av dem har sitt bestemte oppdrag. Algoritmisk informasjonsteori (AIT) ble utviklet av Kolmogorov og Chaitin som en metode til å behandle kompleksitet, spesielt informasjonsinnholdet eller kompleksiteten til en spesifisert sekvens, ved å se hvor omfattende algoritme som trengs for å generere sekvensen.

Ifølge AIT er informasjonsmengden i X (der X kan være strengen av binære tegn, eller en steng med vanlige tegn eller bokstaver i et alfabet) er antall H(X) bits i det korteste programmet som kan generere X.



La oss betrakte en annen streng som er fremkommet ved at en apekatt har lekt med tastaturet på en datamaskin: Mtl3#8HJ;LielSn?ød*nilS … Og anta at denne strengen også er 6 milliarder tegn lang, det vil si at den har samme lengde som strengene vi tok for oss ovenfor. Det er klart at denne strengen i prinsippet er tilfeldig. Ethvert program som skulle skrive ut strengen ville bli omtrent av samme lengde som strengen selv. Det betyr at strengen er algoritmisk ukomprimerbar. Algoritmisk ukomprimerbarhet er en meget god målestokk for tilfeldighet. Strengen er også maksimalt kompleks når vi bruker våre kriterier for kompleksitet.

La oss deretter betrakte en tredje streng med de første 6 milliarder bokstavene i bøkene som vi kan finne i hyllene i et bibliotek. Selv om vi kunne oppnå litt algoritmisk kompresjon, vil den være ubetydelig sammenliknet med lengden av strengen. I praksis vil strengen være like algoritmisk ukompressibel som den andre strengen ovenfor (sett fra et matematisk synspunkt er den tilfeldig). Ut fra den samme betraktning er den meget kompleks. Likevel er kompleksiteten av en annen art enn den vi hadde i den strengen som apekatten laget. For den hadde ingen mening som vi kunne lese. Den tredje strengen inneholder semantisk informasjon. Vi kan forstå meningen med ordene i bøkene. Og grunnen til at den tredje strengen har mening for oss er at vi på *uavhengig måte* har lært språket slik at vi gjenkjenner ordene som er formet av bokstavene i strengen. En slik streng er ikke bare kompleks,

den er uttrykk for det vi kaller *spesifisert kompleksitet*. Dette uttrykket ble første gang brukt av Leslie Orel i hans bok *The Origins of Life* og også av Paul Davies i *The Fifth Miracle*, men ikke i presis betydning noen av stedene. Men spesifisert kompleksitet er blitt undersøkt på en grundig måte av matematikeren William Dembski i *The Design Inference: Eliminating Chance through Small Probabilities*.

Det viktige poenget her er at DNA-sekvensen som koder for et funksjonelt protein *på en og samme tid* er uttrykk for den spesifiserte kompleksitet som er nødvendig for den for å kode for dette proteinet. Den er derfor er algoritmisk ukomprimerbar, noen som innebærer at den fra et matematisk synspunkt er tilfeldig. Paul Davies skriver: "Kan spesifisert tilfeldighet være et garantert resultat av en deterministisk, mekanisk, lovliknende prosess, som kan tenkes i en ursuppe som er overlatt til fysiske og kjemiske lover? Nei, det er umulig. Ingen kjent naturlov kunne frembringe dette."[2] Et annet sted skriver han: "Vi må konkludere at biologisk relevante makromolekyler samtidig innholder to vitale egenskaper: tilfeldighet og ekstrem spesifikasjon. En kaotisk prosess kan kanskje ha mulighet til å frembringe den første egenskapen, men ha neglisjerbar sannsynlighet for å frembringe den andre."

## Kan informasjon "samles opp"?

Vårt spørsmål blir nå: Finnes det noe vitenskapelig indikasjon på at informasjon blir bevart i noen meningsfull betydning av ordet? Hvis svaret viser seg å være positivt, da kan forskningen som gjelder livets opprinnelse spare mye verdifull tid og anstrengelse ved å gi opp å finne en informasjonsteoretisk ekvivalens til en perpetuum mobile maskin.

Vi bør merke oss at det ikke lenger er adekvat å opponere mot å bruke et maskinspråk når vi refererer til organismer. Som vi har sett flere ganger blir et maskinspråk allment brukt i molekylærbiologien av den enkle grunn at proteiner, flageller, celler osv *er* molekylære maskiner. De kan godt være mer enn maskiner. Men så lenge vi taler om deres kapasitet til å utføre informasjonsbehandling, er de virkelig maskiner (utstyrt med software).

Dette medfører noe som allerede er anvendt vitenskapelig på svært mange ulike måter de siste årene, at biologiske maskiner er gjenstand for matematiske analyser generelt og informasjonsteoretiske analyser spesielt. Det er til denne analysen vi vender oss for å få innsikt til å besvare spørsmålet om de molekylære maskinene (uansett type) kan generere ny informasjon. Leonard Brillouin uttrykker i sitt klassiske arbeid om informasjonsteori ingen tvil om hva svaret er. Han sier: "En maskin skaper ingen ny informasjon, men utfører en meget verdifull transformasjon av allerede kjent informasjon."[3]

Tjue år senere er det ingen ringere enn nobelprisvinneren Peter Medawar som skriver: »Ingen prosess som styres av et logisk resonnement – ingen ren tanke- eller en computerstyrt operasjon – kan øke informasjonsinnholdet av de aksiomer og premisser eller observasjonsutsagn hvorfra den stammer.«[4] Av denne observasjon utleder han så at det må finnes en form for lovmessighet knyttet til informasjonsoppsamling. Medawar forsøkte ikke å påvise en slik lovmessighet. Han var fornøyd med å utfordre sine lesere »til å finne en logisk handling som vil kunne supplere informasjonsinnholdet av en hvilken som helst ytring«. Han kom dog med et matematisk eksempel for å illustrere hva han mente. Han påpekte at Euklids berømte geometriske teoremer simpelthen er en måte å »skjære ut i papp den informasjon som allerede ligger i aksiomer og postulater – eller å frigjøre den«. Det er nå engang slik, tilføyer han, at filosofer og logikere siden Bacons tid ikke har hatt problemer med å erkjenne at deduseringsprosessen kun presiserer de opplysninger som allerede foreligger; den skaper ikke ny informasjon overhode.

I den senere tid har William Dembski argumentert for en ikke-deterministisk lov for oppsamling av informasjon, i den betydning at selv om naturlige prosesser som kun involverer tilfeldighet og nødvendighet, kan overføre kompleks spesifisert informasjon, kan de ikke generere denne informasjon.[5]

Der ligger stadig mye interessant og vanskelig arbeid foran oss innenfor dette lovende fagfeltet. Men ikke desto mindre er vi allerede nå i stand til å prøve ut disse ideer på simuleringer som har med livets opprinnelse å gjøre. For hvis informasjon i en eller anden forstand kan bli samlet opp, så må vi av dette logisk forvente at enhver simulering av livets opprinnelse som hevder å få informasjon "gratis" vha. helt naturlige prosesser, på en eller annen måte, får smuglet informasjon inn utefra, på tross av deres påstand om det motsatte. Så hvis vi kan fastslå det siste, har vi i det minste et plausibelt argument for at et informasjons-input er nødvendig for livets opprinnelse. ■

## Referanser og noter

1. *Semantikk* er avledet fra det greske ordet for tegn, og *semiotikk* er teorien for tegn.
2. Paul Davies (1998). *The Fifth Miracle*. Pingvin Press, side 88.
3. Leonard Brillouin (1962). *Science and Information Theory*. 2nd Ed, Academic Press.
4. Peter Medawar (1984). *The Limits of Science*. Oxford University Press, side 79.
5. William Dembski (1997). Intelligent Design as a Theory of communication. *Perspectives on Science and Christian Faith*, 49, 3, side 180-190. Se også hans bok No Free Lunch (2002).